Doctoral Programme in Economics

UNIVERSITÀ DEGLI STUDI FIRENZE

UNIVERSITÀ DI SIENA

UNIVERSITÀ DI PISA

Cycle XXVIII

Department of Economics and Statistics, University of Siena

Director of the Doctoral School: Prof. Ugo PAGANO

Dissertation title:

COOPERATION, RULES, AND MOTIVATIONS

ESSAYS ON MORALITY IN DECISION-MAKING, ECONOMICS, AND INSTITUTIONAL EVOLUTION

Scientific Disciplinary Sector: SECS-P/01

Ph.D. Candidate: Pietro GUARNIERI

Supervisor: Prof. Simone D'ALESSANDRO

ACADEMIC YEAR: 2016-2017

Contents

Introduction: notes on moral motivations in decision-making and institutional evolution **6**

1 Deliberation, discussion, and the intentionality of other-regarding and pro-social decisions **17**

1.1 Introduction 17

1.2 The limits of intuitionism 20

1.3 Epistemic conditions and normative frames 24

1.4 Awareness and the intentionality of other-regarding decisions 27

1.5 Responsibility and the intentionality of pro-social decisions 32

1.6 Conclusions 35

2 Normativity, commitment, and interactive intentionality **37**

2.1 Introduction 37

2.2 The paradox of commitment 42

2.3 The incompleteness of deontology 46

2.4 Variety and conflict in status functions 50

2.5 Institutional formation through *interactive intentionality* 54

2.6 Conclusions 59

3 The effect of ethics meetings on risk-taking behaviour: an experiment **62**

3.1 Introduction 62

3.2 Literature review 66

3.3 Experimental design 69
3.3.1 Hypotheses 73
3.4 Results 74
3.4.1 Univariate 74
3.4.2 Multivariate 76
3.5 Conclusions 79
Appendix: Second phase instructions 81

References **82**

List of Figures

1 The feedback of moral motivations 11

1.1 Awareness as values change 29

1.2 Awareness as ends change 30

1.3 The emergence of responsibility 33

2.1 Conflict as non compatibility of values and of ends 54

2.2 Conflict resolution via emergence of a compatible value 57

2.3 Conflict resolution via emergence of a compatible end 58

2.4 Formation and recognition of new purposes and powers 59

List of Tables

3.1 Social Preference Choice 70
3.2 Rolling the Die: Left or Right Choice 71
3.3 Second Phase: Treatment Overview 72
3.4 Choice side across treatments 75
3.5 Beliefs about Choice Side across treatments 77
3.6 Logit: Choice Left 80

Introduction: notes on moral motivations in decision-making and institutional evolution

Morality is the central concept of the three papers contained in this thesis. Generally speaking, morality refers to "principles concerning the distinction between right and wrong or good and bad behaviours" (Oxford English Dictionary). Here it is taken in a broad epistemological sense. The main hypothesis is that reflection and decisions concerning moral principles constitute a cognitive framework through which subjects know and interact with reality, in particular social reality. Contemporary economics has its own epistemology of morality. In this introduction, I firstly, present some aspects of this epistemology by taking into account how and to what extent economics is able to explain "moral motivations" as determinants of other-regarding, pro-social, and norm-compliant behaviours[1]. The representation of moral motivations that emerges from economics' treatment of cooperation, rules, and motivations is an *ex post* representation. The second part of this introduction illustrates the structure and the objectives of the thesis, while outlining some notes on how to complement economics epistemology of morality with an *ex ante* perspective on moral motivations.

The notion of "social" or "other-regarding" preferences have been admitted within economics' decision-theory paradigm with the intent to account for altruistic and pro-socially oriented decisions by rational individuals. The basic intuition is that in maximizing utility the individual takes into account others' payoffs as a variable of her utility function. Social preferences can take various forms, as shown by numbers of economics

[1] I neither discuss how particular moral-philosophical reflection on morality influence both positive and normative economics (Hausman and McPherson, 1993; Hausman et al., 2016), nor the contribution that economics, and in particular game theory, provide to ethics (Binmore, 1994, 1998).

experiments, in which subjects reveal preferences for altruism (Levine, 1998), overall welfare (Charness and Rabin, 2002), fairness (Rabin, 1993), inequality aversion (Fehr and Schmidt, 1999). These results contributed to the economics' epistemology of morality by demonstrating that agents are not only motivated by self-interest – i.e. self-regarding preferences defined over outcomes – but there exist other-regarding motives such that outcomes concerning others are taken into account in decision.

Experimental literature highlights another type of social preferences, "strong reciprocity", which accounts not only for other-regarding, but also for norm-compliant behaviours. Strong reciprocity has to be distinguished from other types of reciprocity – "direct reciprocity" (or "reciprocal altruism"), "indirect reciprocity" and "network reciprocity" (Nowak, 2006) – which explains the emergence of cooperative behaviours in the context of repeated interactions. Cooperative decisions, in these cases, may be motivated by the self-interested expectation of being reciprocated in the future. The mutualistic nature of these types of reciprocity makes them unable to explain genuinely altruistic and cooperative decisions, for which benefiting others does not imply any kind of benefit for the decision-maker or even inflicts a cost on her.

Strong reciprocity explains why people are willing to cooperate even in one-shot and anonymous interactions, when agents are not motivated by the expectation of future rewards. Gintis et al. (2005, pp.8) defined it as "*a predisposition to cooperate with others, and to punish (at a personal cost, if necessary) those who violate the norms of cooperation, even when it is implausible to expect that these costs will be recovered at a later date*". In particular, many experiments have highlighted as people are willing to renounce to their own payoffs in order to punish non-cooperators (Fehr and Gächter, 2000) or those who do not comply to commonly accepted norms (Fehr and Fischbacher, 2004).

Strong reciprocity has been interpreted as the result of group selection[2]: groups hav-

[2]Group selection has been explained both as genes-driven (*genetic group selection*) and as culture-driven (*cultural group selection*). In the first case, genes are the cause of variation and the modality of transmission. In the second case, variations and differential reproduction involves cultural and informational features. Theories of gene-culture co-evolution provide a synthesis by explaining how cultural changes may condition genetic selection. In the case of morality, the adaptation of cooperative and

ing strong reciprocators are favoured in the competition for resources against groups which do not have them. More precisely, the fitness of members of groups that practice strong reciprocity is increased by the higher degree of cooperation with which they outcompete the groups that lack strong reciprocators. Hence, the evolution of strong reciprocity is connected to the emergence of social norms. Under the lens of strong reciprocity morality is seen as a key evolutionary adaptation, but it is reduced to the efficiency advantages that it provides to the groups which possess it. In general, evolutionary game theory treatment of social norm identifies norm-compliance in terms of stable equilibria in a population. i.e. as the product of interaction among agents whose reciprocal "best-responses" give rise to the spreading of behavioural rules which reveal themselves as fit in the given environment. This implies an epistemology for which individuals are motivated towards norm compliance because they can infer, from the stability of the social norms, the expectation that other individuals will conform to the norm itself.

A peculiar characteristic of "strong reciprocity" motives is that it is not only referred to outcomes involving others, but also to others' intentions to cooperate (Falk et al., 2003)[3]. This "process regarding" feature of strong reciprocity highlights a non-consequentialistic, non-instrumental aspect of moral motivations which has been interpreted in terms of "intrinsic motivations" in the "motivation crowding" literature (Deci, 1971; Bowles, 2008; Frey, 1997; Frey and Jegen, 2001).

"Intrinsic motivations" are defined as "the doing of an activity for its inherent satisfaction rather than for separable consequences" (Ryan and Deci, 2000, p. 8). As such, they are opposed to extrinsic motivations which rest on some external notion of ma-

reciprocity behaviours could at first be sustained by genetic mechanisms, but with the cultural setting of moral codes and norms a new environment for selection and adaptation appeared. For a survey of theories of the evolution of morality see Hodgson (2012).

[3]These intentions-referred preferences are interpreted as *process-regarding preferences* defined as "an evaluation based on the reasons why a state occurred rather than any intrinsic characteristic of the state". The idea is that the process that produces the outcome reveals the intentions of others who have taken part into the process itself. For example, in the ultimatum game, subjects are more likely to reject offers generated randomly by a computer that offers from real participants (Blount, 1995). Moreover, concern for processes provides cues about the socially appropriate behaviour in the decision context (Bowles, 2004, p. 109).

terial reward or punishment. Numbers of studies have proved that the introduction of extrinsic rewards or punishments have an effect on intrinsic motivations (Bowles and Polania-Reyes, 2012). The influence can be positive in the case of *crowding in* – with incentives and intrinsic motivations as complements - or negative in the case of *crowding out* – with incentive and intrinsic motivations as substitutes. According to this literature, the evidence demonstrates that moral intrinsic motivations and instrumental self-interested motivations are "non-separable" (non-additive). This is due to the fact that incentives not only have an effect on instrumental preferences over material outcomes, but also an effect on the salience of the ethical, intrinsically pro-social action, so that the overall effect is non-linear and moral motivation may also be undermined (or over-reinforced). Consequently, it must be acknowledged that economic incentives and institutions have played a crucial role in (social) preferences' change, which for this reason has to be consider endogenous (Bowles, 1998).

The survey sketched above highlights the existence of three approaches, centred respectively on "social preferences", "social norms", and "intrinsic motivations". Concerning the epistemology of moral motivations, one common feature emerges and it regards the circumstance that moral motivations are identified by the three approaches only *ex-post*.

Consistently with rational choice theory, "social preferences" are identified only after the relative ordering has been expressed by actual choices. In other words, the motive behind acting in an other-regarding or pro-social way can be identified as such only after that preference has been expressed or that choice made. This results in an ex-post explanation to the extent that within the standard paradigm it is not possible to identify the moral motivation, unless it has been revealed by an actual and well-defined preference.

Also the conception of the evolutionary emergence of "social norms" entails an *ex post* characterization of moral motivations, to the extent that it equates norm-compliant decisions to behavioural traits. The evolutionary explanation of norm-compliance is available only after that norm has been established, through evolutionary processes,

as a stably followed and self-enforcing behavioural rule. Those evolutionary processes accounts for those behaviours only after they have persisted as a stable institution in the given environment. However, this explanation can be given only after the relevant process of evolutionary emergence is completed, and in any case does not specify the norm content and why people are actually committed to follow (or not to follow) it.

Motivation crowding theory characterizes moral motivations only *ex post*, in the sense that it can provide a description of them only after the introduction of economic material incentives. It is assumed that moral motivations exist as intrinsic motives, but they can be analytically identified only by comparing them with economic, extrinsic motivational factors. What interests economic analysis is the effect of the introduction of economic incentives on intrinsic motivations taken as given. It is an *ex post* characterization of moral motivations because intrinsic motivations are described only in terms of effects, i.e. as the consequences that the introduction of material incentives has on other-regarding or pro-social behaviours.

These considerations are not meant to deny the relevance of the *ex post* aspect of moral motivations that can be deduced from economics' explanation of other-regarding, pro-social and norm-compliant decisions. The point is that these explanations reveal only one aspect of moral motivations and of their possible impact on decisions in socio-economic-institutional contexts. Consequently, three questions motivate the present research: a) what does the *ex-post* characterization of moral motivations lack?; b) how could it be represented as an *ex ante* motivational factor?; c) does this *ex ante* motivational factor influence other-regarding, pro-social, norm-compliant decisions, and hence institutional interplay?

The main hypothesis that is investigated here is that moral motivations can feed back on the process of social preference formation, on the collective definition of social norms, and on the definition of what counts as an economic incentive in given social and historical contexts. In order to complement the *ex post* epistemology of morality, I investigate under which epistemic conditions moral motivations can act as an *ex-ante* determinant of preferences and behaviours, such that pro-sociality, normative commit-

ment and the effect of economic incentives are influenced by them. These objective let me specify three sub-hypothesis represented as feedback loops in figure 1

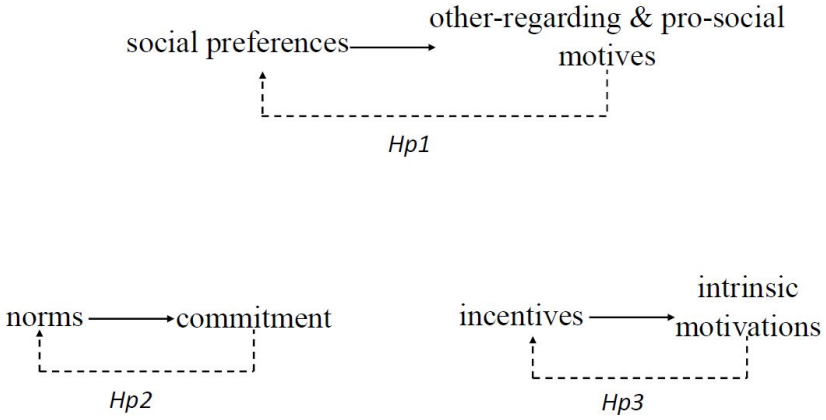

Figure 1: The feedback of moral motivations

The *ex post* epistemology of morality connected to the economics' treatment of social preferences, norms and intrinsic motivations is only able to develop an explanation which goes from left to right: from social preferences to pro-social and other regarding motives, from evolutionary stable institutional norm to norm commitment, from the effect of economic incentives to intrinsic motives. I investigate whether moral motivational factors, as specified on the right-hand side, can have a feedback influence on the decision and evolutionary factors on the left-hand side. More precisely, the present research analyses the epistemic conditions under which the above hypothesized feedback are effective in influencing economic decisions and institutional evolution.

In order to try to characterize from the *ex-ante* perspective the role of morality in decision-making and institutional evolution, the thesis discusses recent attempts to integrate theories and findings in moral and social psychology within economics' explanation a) of other-regarding and pro-social decisions, and b) of normative institutions and rule-following.

As far as other-regarding and pro-social decisions are concerned, the introduction of Dual Process Theory among the conceptual and operational tools of experimental

and behavioural economics (Alos-Ferrer and Strack, 2014) has brought about a great emphasis on the role of *intuition* as a driver of them. Experimental studies have proved that intuitive reactions favour cooperative responses in social dilemmas, confirming the so-called "social-heuristics hypothesis" (Rand et al., 2014; Rand, 2016). Intuition, accompanied by emotions, has been considered at the basis of deontological judgments, with utilitarian rationality attributed to controlled (slower) processes and to the role of meta-morality (Greene, 2007). This evidence has been referred to frame evolutionary interpretations where intuitions and moral emotions lead to evolutionary advantages in the context of competition between groups by fostering cooperation within groups (Greene, 2013).

The objective pursued in the first chapter of the thesis is to show that morality is not reducible to these non-reflective decision-making modes, and always presupposes a (parallel or previous) reflective mode. In order to do this, I take into consideration the possibility that reflection is a process which frames the epistemic contexts in which other-regarding and pro-social decisions are made, setting the (normative) relevance and salience criteria which identify other-regarding and pro-social options as such – and eventually makes the implementation of social heuristics possible. This framing process, by identifying which actions and which consequences of action are contextually relevant and salient for the identification of other-regarding and pro-social options, can also be viewed as grounding social preferences formation. Thus, *deliberation* and *discussion* are interpreted as the interactive modes which develop the reflection process.

As far as normative institutions and rule-following are concerned, the reference to *collective intentionality* (Gallotti and Frith, 2013) provided with a psychological ground theories of institutions based on rules and commitment (Searle, 2010; Gilbert, 2013). In this literature, collective intentionality is viewed as the epistemic basis of the recognition that makes institutional entities exist as accepted and adhered social functions. This approach challenges the standard economics' explanation of institutions as based on equilibria and *self-enforcement*. Consequently, a problem of "unification" emerges and it requires the development of an accepted conception of institutions capable to overcome

the rule-equilibrium dualism (Aoki, 2011; Guala and Hindriks, 2014).

The objective pursued in the second chapter of the thesis is to rethink conceptions of normative commitment, by reinterpreting collective intentionality as a form of interaction. This can be done by allowing for multiplicity of alternative "powers" (potentiality of action) and "purposes" (the aim of the action) which identifies institutions as "status functions" in Searle's social ontology. Under the assumption of multiplicity of status functions, collective intentionality can be reinterpreted as the result of an interactive process of selection of values and ends defining those institutional powers and purposes. This interactive process, which I name "interactive intentionality" is essentially moral in so far as it activates practical reasoning in an inter-subjective mode. It can be viewed as a framing process that defines which actions and which consequences are normatively right (valuable and to be pursued) in the institutional context, so grounding the epistemic definitions of extrinsic enforcement (punishments and rewards), while engaging people in intrinsic commitment.

An attempt to define the *ex-ante* role played by *practical reason* in grounding the epistemology of other-regarding, pro-social and norm-compliant decisions is common to the two objectives highlighted above. This attempt is based on the idea that decisions concerning values and ends act as a prior, setting the general epistemic framework under which both means and ends are identified as normatively relevant actions and consequences. My interpretation of practical reasoning has various sources of inspiration. It relies on conceptions on the role of *frames* in the interpretation of decision situations, developed in the context of theories of team-reasoning (Gold, 2012; Karpus and Gold, 2016). It derives its characterization of the relationship between values and ends from the philosophical discussion concerning "incommensurability" and "incomparability"(Chang, 1997). It re-elaborates this relationship applying the notion of "intentionality" in order to analyse how the processes of selection concerning values and ends structures the relation between the decision-maker and the world of options – both actions *per se* and consequences – to which values and purposes are attached. Reflecting on collective intentionality, it highlights how these processes are essentially inter-subjective

end evolutionary(Gold and Sugden, 2007; Tomasello et al., 2005; Tomasello, 2009, 2010).

The result is a characterization of morality as emerging from deliberative and interactive practical reasoning. This characterization goes together with the recognition of the "discoursive" nature of ethics (Habermas, 1990), which in turn highlights a linguistic presupposition of deontological normativity. Normative definitions are constituted in the context of linguistic interactions through the performance of ethical discussion. Habermas summarizes this conception in the *discourse principle*: "only those norms can claim to be valid that meet (or could meet) with the approval of all affected in their capacity as participants in a practical discourse" (Habermas, 1990, p. 197). Interactive argumentative procedures lead to agreements on deontological references which orient individuals' decisions to the extent that they are rationally elaborated and shared. The deliberative and linguistic dimension of morality is emphasized in this passage from Geoffrey Hodgson's *From Pleasure Machines to Moral Communities*:

> "the evolution of morality combined moral feelings that evolved among our ape-like ancestors with discursive capacities that are much more recent. [...] notwithstanding its genetic foundation, morality and its evolution have to be understood also as social phenomena. Morality is much more than collections of individuals with moral or cooperative attributes, and typical models of the evolution of altruism or cooperation are inadequate in their treatment of morality. Reference to moral systems is required to explain the persistence of moral or cooperative attributes, especially when we acknowledge that humans can resist emotions and reflect on their interests." (Hodgson, 2012, pp. 118-119)

By dealing with reflection and interactive intentionality the thesis tries to grasp some distinctive aspects of the epistemology of moral motivations and of normativity. The epistemological point of view favours the theoretical understanding of conditions which play behind social preferences and norms, affecting their formation and emergence. The acknowledgment of these conditions entails a conception of the decision-maker as

interacting with the epistemic and social context of the decision. This interaction is measured by the *awareness* concerning the values through which actions are evaluated and the *responsibility* towards the ends pursued to obtain specific consequences.

Interaction between individuals via deliberation and discussion, as modes of reflection and of interactive intentionality, can be viewed as the way in which the decision-maker gets aware of and responsible for normatively relevant actions and salient consequences in the decision situation. The effects of these reflective and interactive processes on social preferences and on "moral" decisions are tested in the third chapter of the thesis through an experimental design which investigates the effects of ethic deliberation and discussion on risk-taking for other. The results confirm the opportunity to develop an *ex ante* conception of moral motivations, not limited to what is *ex-post* revealed by other-regarding or norm-compliant behaviours, nor to the effects of extrinsic economic incentives on them. The thesis fosters this *ex ante* perspective by showing that moral motivations are also formed through interaction in practical reasoning. They frame the way in which the decision-maker thinks at the decision itself, becoming aware and responsible for its normative implications.

To sum up, the three papers which constitute the thesis proceed as follows.

The first paper focuses on the epistemic conditions of other-regarding and pro-social decisions. These conditions consist in relevance and salience criteria concerning possible actions and consequences of decisions available in the decision context. Under the lens of *intentionality*, it becomes possible to describe processes of identification of normative references (values and ends) and social determinants (the boundaries of the relevant social community) which set the epistemic necessary conditions of other-regarding and pro-social decisions. From this perspective, *deliberation* - interpreted as self-reflection - and *discussion* – interpreted as social reflection – play a fundamental role, carrying out the framing function which makes the interpretation of other-regarding and pro-social options available in the decision situation, while enhancing *awareness* and *responsibility*. This in turn offers the opportunity to reintegrate recent "intuitionist" tendencies in moral psychology and experimental economics within a unified paradigm in which moral

reflection and reasoning define the normative cues which make intuitions applicable in specific decision contexts.

The second paper discusses Searle's description of social ontology in terms of deontological constitutive rules and collective recognition. It aims at contributing to the "unification" debate, by integrating Searlian conception of constitutive normativity with an epistemology of rule-following capable to illustrate processes of *formation* of institutions. Social ontology *per se* cannot account for the emergence of institutions from *alternative* normative definitions of status functions. I make the hypothesis of *interactive intentionality* as an inter-subjective mode of practical reasoning concerning the selection of values and ends. This interactive process of selection frames relevance and salience criteria which represent the epistemic basis of status functions' identifiability and of normative commitment.

The third paper[4] is based on an experimental setting and investigates the effect of bank ethics meetings on individual risk-taking behaviour when there are monetary conflicts of interest between the decision-maker and the client. In the decision setting, participants are either assigned to investor or client roles. Investors choose between two risky gambles, one of which always implies a better outcome for the investor but exposes the client to higher risk. In the main treatments (i.e. Ethics meeting), in contrast to the Baseline treatment, participants also discuss – before their investment decision – the consequences of their choice within a group of decision-makers (i.e. peers). The results show that, in treatments with ethics meetings, investors tend to choose more often the gamble less risky for clients, so confirming a pro-social effect of moral deliberation and discussion.

[4]In collaboration with Fracensco Feri (Royal Holloway University of London) and Caterina Giannetti (University of Pisa)

CHAPTER 1

Deliberation, discussion, and the intentionality of other-regarding and pro-social decisions

1.1 Introduction

This paper discusses the role of reflection and moral reasoning in other-regarding (OR, henceforth) and pro-social (PS, henceforth) decisions, by adopting the point of view of *intentionality*. "Intentionality" is a technical notion in philosophy. Stanford Encyclopedia of Philosophy defines it as "the power of minds to be about, to represent, or to stand for, things, properties and states of affairs" (Jacob, 2014). Hence, we can consider the intentionality of OR and PS decisions as defined by the mode of reference to the "object" towards which they are directed. The intentionality of OR decisions is defined by the reference to *the benefit of other(s)* – one or more individuals addressed by the OR decision. The intentionality of PS decisions is defined by the reference to *the benefit of the society as a whole* – i.e. the generality of people included in the definition of some relevant group. Aim of the paper is to show that OR and PS decisions refer to the respective objective through a normative lens.

The analysis of these forms of intentionality reveals context-dependency conditions which are not taken into account in the other-regarding preferences and (pro)social norms literature. These conditions are epistemic to the extent that they concern the

decision-maker's beliefs about the decision itself. I focus on the epistemic conditions of OR and PS decisions in order to show that they are determined by reflection and moral reasoning. Accordingly, I investigate deliberation and discussion as frames of OR and PS decisions.

The intentionality of OR and PS decisions entails a *mode* and a *content*. The mode concerns the action which is considered benefitting others or the society. The content concerns the state of the world which is realized as the consequence of the decision to benefit others or the society. I make the hypothesis that modes are specified by normative *values*, while contents are specified by normative *ends*. Values and ends define the epistemic conditions which make OR and PS actions and consequences identifiable in the decision context. Thus, I consider that rational scrutiny and selection of values and ends (i.e. *practical reason*) determine the framing within which OR and PS decisions are taken. Following team reasoning approach (Bacharach, 2006; Gold and Sugden, 2007), I consider that practical reasoning leads the decision-maker to think at the decision in a "collective mode", so that the OR and PS options, available in the decision situation, are identified according to their normative (universal) relevance.

The intentionality perspective fosters a reconsideration of the role of reflection and moral reasoning in OR and PS decisions. This is in conflict with recent trends in moral psychology and experimental economics literature which support the idea that intuition plays a primary role in cooperative and moral decisions. In the last years, an "intuitionist" paradigm, based on Dual Process Theory (Evans, 2008), oriented research in moral psychology and led to collect evidence proving that moral judgment is driven by intuitive responses (Haidt, 2001), that deontological thinking is sustained by affective and un-reflective reactions (Greene et al., 2001), that intuitive heuristics favours cooperation in social dilemmas (Rand, 2016).

However, the overall evidence that intuition triggers cooperation and morality are challenged both on the methodological and empirical ground[1]. It is hence useful to

[1]See Umphress et al., 1997; Suter and Hertwig, 2011; Kahane et al., 2012; Tinghög et al., 2013; Krajbich et al., 2015; Achtziger et al., 2016; Tinghög et al., 2016.

further investigate the role of deliberation and moral rationality in OR and PS decisions. This is a necessity which dual process theorists are aware of when they admit that "little is known about the psychology of moral reasoning and that it may yet prove to be a potent social force" (Paxton and Greene, 2010). Also within the intuition based paradigm it is recognized a connection between reflection and normativity, as witnessed by the following definition of "moral reasoning"[2]:

> "Conscious mental activity through which one evaluates a moral judgment for its (in)consistency with other moral commitments, where these commitments are to one or more moral principles and (in some cases) particular moral judgments". (Paxton and Greene, 2010).

Other threads of experimental research account for the role of normative values and norms as motivating OR an PR decisions (Liu and Ditto, 2013; Ditto et al., 2009; Schram and Charness, 2015). There is evidence that moral decisions are connected with longer deliberation time in circumstances in which the ethical "right" is salient and opposed to "wrong" options (Gunia et al., 2012; Shalvi et al., 2012). But this connection appears non-linear, and the functioning of moral deliberation appears complex and case-dependent (Ham and van den Bos, 2010; Zhong and Liljenquist, 2006; Moore and Tenbrunsel, 2014)

This paper further investigates the relation between moral reasoning and normativity by discussing the epistemic conditions behind the intentionality of OR and PS decisions. By presenting a simple model of the effects of moral reflection about values and ends on the decision-maker beliefs, it proposes a possible reconciliation of intuition and deliberation. This implies recognizing that moral reflection has a specific framing role, acting on the perception of the context of decision and hence providing moral intuition with defined references.

Social heuristics and moral intuitions are conditioned by reflective and interactive processes which, by constituting normative frames, provide OR and PS decisions with

[2]See Nichols and Mallon, 2006; Harman et al., 2011; Paxton et al., 2012

an interpretation of relevant options in the decision situation. Social heuristics and moral intuitions are subject to specific epistemic conditions which I deal with in terms of "context-dependency" of OR and PS decisions. The analysis reveals the limits of validity of the intuitionist hypothesis and points out the necessity to re-position it within a conception of moral *deliberation* and *discussion* as processes through which the normative relevance of decisions are framed and (re)cognized. I analyse deliberation as "self-reflection" – i.e. self assessment of one's own normative values and ends – and discussion as "social reflection", i.e. as an inter-subjective and interactive assessment of values and ends[3].

In section 1.2, I present intuitionist paradigm and the limits which concern its applicability to the generality of OR and PS decisions. Section 1.3 shows that the limits of applicability of intuitionism coincide with defined epistemic conditions, specified as normative criteria by the intentionality of OR and PS decisions. I present some distinctive features of reflection-based normative framing in sections 1.4 and 1.5, which respectively highlight the role of *awareness* and of *responsibility* in OR and PS decisions. Common to awareness and responsibility is the *reflection* process consisting in the decision-maker thinking-back (reflecting, in this sense) on 1) her own criteria of decisions 2) the weight of her decision with respect to the cognition of the social context. Deliberation and discussion, by activating this re-thinking, put the decision-maker in the condition to attune her decision to the relevant normative references in the decision situation. Section 1.6 summarises and concludes.

1.2 The limits of intuitionism

The discussion concerning the role of intuition and reflection in OR and PS decisions is articulated in the literature within the framework of Dual Process Theory. Dual Process Theory explains decision-making as the result of two distinct processes: one

[3]The reference to discussion goes together with the recognition of the role of communication in other-regarding and pro-social decisions. It is inevitably downplayed by intuitionism, in spite of several experimental studies in economics proving that communication sustains cooperation in social dilemmas (Balliet, 2010).

intuitive, automatic, effortless, and often accompanied by emotions; the other deliberative, controlled, slow, effortful, associated with reasoning (Evans, 2008; Kahneman, 2011; Alos-Ferrer and Strack, 2014). Dual Process Theory is at the basis of what can be considered a real "intuitionistic turn" in the psychology of moral judgment. Among the contributions which started the turn in the early 2000s, Jonathan Haidt proposed the "social intuitionist model" (Haidt, 2001), challenging the idea, till then dominant, that moral judgement is based on reasoning (see Piaget, 1965; Kohlberg, 1969).

The social intuitionist model is based on two basic claims: 1) "moral judgement appears in consciousness automatically and effortlessly as a result of moral intuition"; 2) "moral reasoning is an effortful process, engaged *after* a moral judgment is made, in which a person searches for arguments that will support an already-made judgment" (Haidt, 2001, pp-818-19, emphasis added). In other words, reasoning comes second, after intuition and after judgment itself. Moreover, reasoning serves two further ancillary functions: a) it has a communicative (social) function, supporting self-justification and persuasion of others; b) it has a self-critical function, reviewing the initial intuitive reaction and judgment, so that they can be overridden by logical reflection or by arguments in favour of another counteracting intuition (Haidt, 2001, 2003, 2007).

On the same line of research, Joshua Greene and colleagues found evidences in favour of the idea that deontological/anti-consequentialistic judgements (judgements referred to the respect of duties and rights) are driven by affective, automatic reactions, while utilitarian/consequentialistic judgements (judgements concerning the "greatest good for the greatest number") are associated to deliberative/controlled responses (Greene et al., 2001, 2004; Greene, 2009). Differently from the Haidt's intuitionsit model, the two modes of moral thinking have a proper and ubiquitous (parallel and possibly conflicting) role in moral judgments, even if, also in this theoretical framework, emotions are still primary as determinants of moral deontological responses, with utilitarian rationality mainly attributed with the role of meta-morality (Greene, 2007, 2013).

The "Social Heuristics Hypothesis" (SHH, henceforth) maintains that "pure cooperation" – i.e. costly cooperative decisions not motivated by individual advantages –

is triggered by intuition (Rand et al., 2012, 2014; Rand, 2016). Several experiments have proved that in the context of social dilemma[4] fast decisions are relatively more cooperative. On the contrary, deliberation tends to favour more selfish decisions. In the framework of SHH deliberation is viewed as grounding individual's payoffs maximization. It sustains cooperation only in strategic contexts in which the cooperative option provides the individual with the expectation of future reward. Deliberation is hence situational and "allows us to adjust to the specific social situation we are facing at any given time, overriding the intuitive response if that response is not actually payoff maximizing in the current setting." (Rand, 2016, p. 2). The SHH is grounded in the idea that successful cooperative behaviours learnt through experiences are internalized as rules (generalized default responses) which are acted automatically (Zaki and Mitchell, 2013). This idea is consistent with evolutionary explanations that interpret intuitive heuristics as fostering cooperation within groups, and hence as determining an adaptive advantage. (Nowak, 2006; Greene, 2013).

Within the intuition-based framework deliberation has a secondary function: it can intervene to revise the intuition-based moral judgment, it acts on the intuitive cooperative decision if not optimal for the individual, but in any case it does not represent an input to the decision process. I consider that the external validity of these intuition-based hypotheses and relative findings is limited to decisions contexts in which the normative identification of moral or cooperative options are given and shared among individuals. Intuition as such depends on the availability in the decision-context of all the relevant information which make the decision-maker recognize the situation as the one in which the intuitive response and the heuristics apply. This condition is not always met, so that we cannot consider the intuitionist model as representative of the generality of OR and PS decisions.

Moreover, within the intuitionist paradigm, it is underestimated the capability of reflection, via deliberation and discussion, to modify the context of decision influencing

[4]One-shot, anonymous cooperation games, with cooperative choice increasing social welfare, and individual incentive to act selfishly, i.e. with optimal selfish choice and the cooperative choice costly to the individual.

the normative identification and evaluation of relevant actions and salient consequences in the decision-situation. What follows is an attempt to demonstrate that moral reasoning have the power to change the epistemic conditions of decision situations, so that individual's beliefs concerning it may change, making the moral intuition or cooperative heuristics not fit anymore. Hence, I focus on the epistemic conditions behind the possibility of moral intuitions and social heuristics in order to trace the limits of applicability of the intuition-based models. I deal with these conditions in terms of *context dependency* of OR and and PS decisions.

Moral intuition and social heuristics are subject to context dependency to the extent that the default response internalized in the heuristics may be cognized as cooperative and moral (or alternatively selfish and non compliant to morality) depending on the actual cognition of the normative meanings attached to the alternatives available in the decision situation. I discuss two types of context-dependencies influencing how the decision-maker views the decision itself, recognizing (or not recognizing) the normative nature of the available options. Context-dependency may affect:

a the beliefs concerning the relevance of normative values and the salience of ends;

b the beliefs concerning the application of norms and rights to relevant subjects.

These two types of context-dependency can be read as deriving from two forms of "bounded rationality". They respectively involve possible decision-maker's ignorance concerning: a) her own normative criteria of evaluation and selection of actions and consequences in the decision situation; b) her cognition of the social context and relationships, which grounds her compliance to norms (only) under determined social conditions. A decision-maker may not acknowledge the relevant criteria of decision identifying which is the OR or the PR option in the decision situation. She could behave in a OR or PS way if and only if those criteria, both concerning normative values/ends and the *subject of normativity*, are given, known, and shared, in the decision context. Unless these criteria are met and are available to the decision-maker intuition cannot recognize the decision situation as the one in which the intuitive moral or social response applies. The

external validity of the intuition-based models is limited by these epistemic conditions.

My hypothesis is that reflection and moral reasoning compensate for the two forms of normative bounded rationality, providing social heuristics and moral intuition with the context of reference that they require to be activated. In this sense, deliberation and discussion are necessary conditions of OR and PS decisions. They frame decision situations and hence they are not only *ex-post* with respect to intuition based judgments and decisions. Deliberation, as *self-reflection* – i.e. self assessment of one's own normative values and ends – has also en *ex ante* role with respect to moral choices and judgments. Discussion as *social reflection* – i.e. as an inter-subjective and interactive assessment of normative values and ends – is not mere *ex-post* justification or persuasion, but the definition of shared normative references which identify the relevant moral community.

1.3 Epistemic conditions and normative frames

The applicability of intuitionism and of SHH is limited to decision contexts that satisfies defined epistemic conditions. In order to trace these limits, I describe two peculiar decision situations, in which any sort of social heuristics or intuition is *per se* unable to specify OR or PS decisions. The two circumstances – connected the two types of context-dependency and bounded rationality presented above (section 1.2) – are respectively affected by lack of relevant knowledge concerning a) *how to normatively interpret actions and consequences* and b) *given social context and relationships defining the subject of cooperative norms.* In the first case, which refers to OR decisions, bounded rational individuals lack pieces of information about the decision context so that social heuristics does not activate. In the second case, which refers to PS decisions, the pre-existence of group boundaries and social ties may generate circumstances of exclusion and conflict, for which the decision-maker may not recognize that acting in accordance to a social heuristics causes disadvantages to someone else or exclude someone else from the benefit of cooperation. In this context, the application of the social heuristic gives rise to anti-social outcomes.

Two epistemic conditions derive from the intentionality of OR and PS decisions. OR decisions rest on the cognition of "the benefit of other(s)", while PS decisions rest on the cognition of the "benefit of the society as a whole". These addressed intentional contents entail that the decision-maker, while opting for the OR or the PS decision, believes that the other(s) or the society as a whole will benefit from her decision (*epistemic condition 1*). Moreover, since the benefit of others depends on the others experiencing it as a benefit, she must believe that her OR or PS decision is recognized as such by the others or in the relevant social context (*epistemic condition 2*). These two epistemic conditions can be translated in conditions of social heuristics and moral deontological intuitions. Let us interpret a social heuristics as a behavioural rule which dictates the decision-maker to do the action Γ in order to get the consequence γ, when she faces the decision situation C^1. This heuristics gives rise to an OR or PS decision to the extent that both the action and its consequences are interpreted by the decision-maker as OR or PS. Moreover, in order to count as an OR or PS decision the decision-maker must believe that also the others referred to in the intentional content of the decision recognize them as an OR or PS action and consequence. These conditions – the circumstance that both the decision-maker and receivers of the decision recognize the OR and PS nature of the adopted decision – are among the conditions defining the features of a decision context C^1. It is a necessary condition of a OR and PS decision that the two epistemic conditions 1 and 2 are simultaneously met. Otherwise, the same action and the same consequences of that action are not anymore identified as OR or PS. Only under these conditions social heuristics leads to OR and PS decisions, and this is not always the case.

As far as OR decisions are concerned, we can notice that the two just mentioned conditions may be affected by decision-maker's bounded-rationality. The decision-maker can believe or not believe to be in the decision context C^1 in which the two conditions hold. Let us consider the possibility that she believes to be in the decision situation characterized by contextual epistemic conditions C^2. She can for example believe that action Γ and its consequence γ do not benefit the other; or she can believe that the other(s) believe(s) that action Γ and its consequence γ do not benefit the other(s) –

i.e. the receiver of the decision. If this is the case, the decision-maker believes she is facing decision situation C^2, while she is actually in the decision situation C^1. As a consequence, the social heuristics would not activate and the OR decision would not be made. I will further analyse these epistemic conditions of OR decisions in terms of *awareness* concerning the decision situation in the next section (section 1.4).

As far as PS decisions are concerned, we can notice that the satisfaction of the two epistemic conditions above depends on the availability of relevant information concerning social context and relationships. Let us suppose that a social heuristics dictates the decision-maker to cooperate with fellow members for the benefit of the society doing Γ and obtaining γ, when he faces the decision situation C^1. The cognition of the boundaries of the relevant social group and ties among fellow members affects the epistemic conditions which characterize the decision situation C^1.

Let us suppose that the society is composed by two groups of people: the ones included in the group of cooperators who have internalized a social heuristics and who benefit from cooperation; the others excluded from cooperation and from its benefits. Considering the society as a whole and so also the group of excluded people, we can consider that in this decision situation social heuristics of included individuals is actually anti-social in so far as action Γ and consequence γ are indeed damaging others and causing social exclusion and conflict. They believe to act pro-socially in C^1, but they are not, in so far as their behaviours cannot be recognized as PS from the point of view of the group of excluded. If the decision-maker would realize that her decision is actually damaging members of the society (according to a cognition of the society in which the excluded are included), the second epistemic condition would be violated, and she would not perceive her decision as PS.

Beliefs concerning the definition of the relevant social group of cooperators affect the identification of actions and consequences as PS. In the described situation only acknowledging the society as a whole (the group of included plus the group of excluded) as within the boundaries of the epistemic relevant social group, would reveal the possibility to act Δ towards the consequence δ: cooperating also with non fellow members and in-

cluding the excluded to the benefit of cooperation. In this sense the identificaiton of the PS option Δ and δ depends on the circumstance that the decision-maker believes to be in decision context C^2 in which the relevant social group includes the whole society. This circumstance calls for the notion of "responsibility" which will be presented in section 1.4 to illustrate how this comprehensive (universal) moral acknowledgment may emerge.

The two presented cases highlights two peculiar senses in which we can say that moral intuition and social heuristics are under-determined with respect to contexts. More precisely, it has been underlined as context dependency is due to decision-maker's beliefs concerning the decision context itself. In this sense, moral intuitions and social heuristics are under-determined with respect to epistemic conditions. My hypothesis is that deliberation and discussion may intervene in these cases, contributing both to the understanding and the change of decision contexts. By acting on the epistemic conditions, deliberation and discussion can influence the formation of social heuristics, preferences, and norms. To account for the framing function linked to the reflection processes involved by OR and PS decisions, it is necessary to further specify their intentionality and to clarify their effect on the epistemic conditions of decisions.

1.4 Awareness and the intentionality of other-regarding decisions

The intentionality of OR and PS decisions involves an evaluation process concerning actions and consequences themselves. This evaluation process forms the intentionality of OR and PS decisions by attributing an intrinsic *value* to the relevant actions and considering the targeted consequence as the *end* which ought to be pursued against alternative others. In this sense, the intentionality of OR and PS decisions is essentially normative. The attribution of values and ends set the normative standard against which actions and consequences are recognized as OR or PS in the decision situation. OR and PS decisions presuppose that a normative selection of relevant actions and salient ends have been already developed, constituting the frame of available OR ad PS options and

setting the conditions of their recognition.

We can rewrite the epistemic conditions of social heuristics and clarify that they presuppose the assignment of normative values and ends to actions and consequences. While choosing OR or PS options, the decision-maker believes that the action he is performing has an intrinsic value which motivates him to get to a specific state of the world which he envisages as an intrinsic end. If she believes so, and if she believes that the others beliefs so, she considers that the action and the consequence of the action are compliant with what she believes to be the benefit the other(s) or of the society as a whole.

To show how the belief concerning the normative characterization of the decision context may condition OR and PS decisions, I introduce the following notation. I refer to values attributed to actions with Greek capital letters and to the ends which are the envisaged consequences of those actions with small Greek letter. The intentionality of pro-social decisions takes hence the form of $\Gamma \rightarrow \gamma$, with the arrow referring to the intentional direction of the OR or PS motivation represented by the value Γ towards the state of the word represented by the end γ.

Considering OR decisions, let us suppose, for example, that the decision-maker in the initial moment t^1 believes to be in the decision context C^1, in which doing $\Gamma \rightarrow \gamma$ is OR, while $\Delta \rightarrow \delta$ is not. For example, she may consider donating a cigarette to her friend as an altruistic act and its consequence (her friend smoking) a benefit to him (the friend). This would mean that in t^1 she beliefs that the decision situation is such that both her and the receiver of the consequences of her decision believe that the donation of the cigarette is "altruistic" in the standard sense of to benefit other at a personal cost.

However, we may consider the possibility that she starts reflecting about the decision situation and after a certain amount of time (at time t^2) she gets to the opposite conclusion concerning the donation decision, doing $\Delta \rightarrow \delta$ – i.e. not donating and not letting her friend smoke – but still interpreting this decision as the altruistic one. What could have happened in the meanwhile? If we exclude preference change, we can consider the possibility that a change in beliefs occurred, influencing the decision-maker cognition of

the decision situation. In particular, I make the hypothesis that two different but related processes may determine a change in decision-maker's beliefs concerning the decision situation itself. At the end of these processes she believes to be in the decision context C^2 in which $\Delta \rightarrow \delta$ is believed as the OR decision: i.e. she believes that not donating is the OR (altruistic) decision and moreover she believes that also the friend believes so. On the contrary, $\Gamma \rightarrow \gamma$ (donating) is not believed anymore as OR. I consider that these processes accounts for the way in which *awareness* grounds OR decisions.

To illustrate the two processes which may lead to such a change in belief through awareness enhancement, let us consider the following figures. Figure 1.1 depicts the first process in which the change in belief about the decision situation is a consequence of a change in values attributed to relevant actions. Figure 1.2 depicts the second process in which the change in beliefs about the decision situation is a consequence of a change in the ends according to which consequences are interpreted and evaluated.

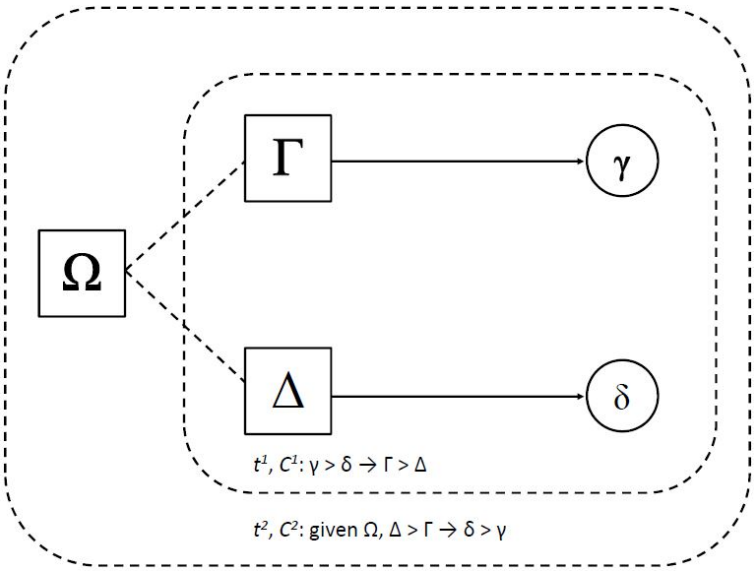

Figure 1.1: Awareness as values change

As far the first type of process is concerned, we have that in the initial situation (t^1, C^1) the decision-maker believes that $\Gamma \rightarrow \gamma$ is OR. Hence, we can consider that she has defined (social) preferences over consequences $\gamma \succ \delta$, and this implies that she

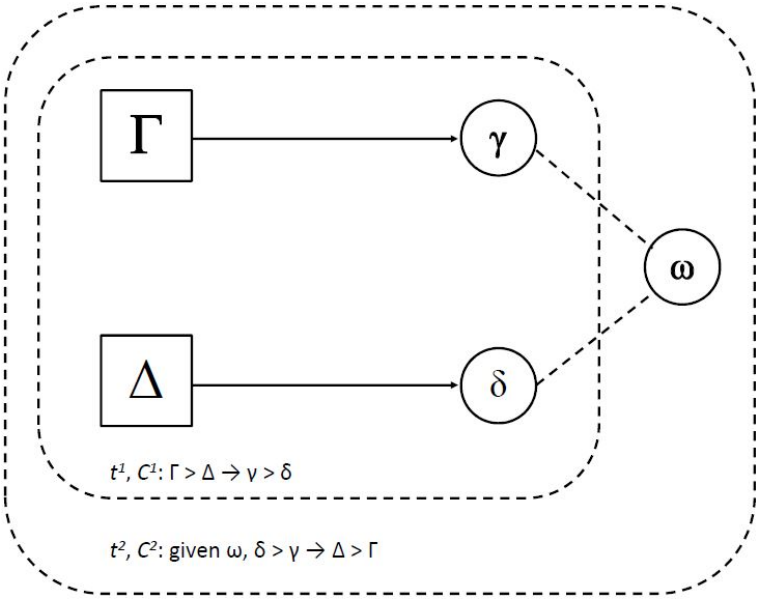

Figure 1.2: Awareness as ends change

has analogous preferences over actions $\Gamma \succ \Delta$. Let us suppose that the decision-maker starts thinking about what is best or more just to do in this situation, and that she develops the awareness that actually "health" is an important intrinsic value to her. The discovered value is represented in figure 1.1 by Ω. It can be considered representative of an alternative action not available in the original set of options (for example, the action "talking to the friend explaining the risks linked to smoking"). She may thus revise her first decision so that in t^2, she prefers action Δ (not donating the cigarette) to action Γ (donating) and so, consequence δ (her friend not smoking) to consequence γ (her friend smoking).

What is important to underline is that it is the whole decision situation at being changed by the acquired awareness (and not only preferences). The decision-maker does not believe anymore that the act of donating is OR, and that the consequence consisting in her friend smoking is benefiting to him. Moreover, she believes that the friend would believe the same if only he could get to the same awareness, recognizing the intrinsic value of "health". The decision has changed from one concerning the value of "being

generous toward a friend" (and hence, from one concerning the act of donating *per se*) to one concerning the value of "protecting your friends' health" (or the intrinsic value of "a healthy life"). The latter normative interpretation appears superior to the former, to the extent that it results from a deeper scrutiny of the available options. A whole translation of what is OR in the decision situation occurs. In t^2, C^2, "being generous to a friend" gets synonymous with "protect his health". Given this change, the decision-maker is acting in an OR way even if now she prefers not to donate.

The change in the evaluation of what is OR happens also in figure 1.2, but in this case it is the envisaged consequences of the decision, as identified by normative ends, at being subject to change after awareness occurs. In this respect, the decision-maker, with time, may reconsider the ends under which the consequences of the decision at stake are evaluated. Thus, she may realize that an alternative end – besides those strictly associated to letting the friend smoke (γ) or not smoke (δ) one of her cigarette – can orient her decision. The decision-maker may for example get aware of the ill effects of smoking so that she may feel obliged to pursue the end of "convincing the friend to give up smoking" (ω). Once she realizes this possibility, she may act in order to obtain the friend quitting as the consequence of her decision. The decision context changes from one in which the end ω is not at stake (C^1 in time t^1) to one in which it is present (C^2 in time t^2). An alternative consequence of decision (identified by ω) emerges and it represents any outcome of decision in which the decision-maker gives up smoking (and avoids the ill consequences of smoking).

The change in the attribution of values and ends can be accompanied by a shift in preferences, with the decision-maker passing from preferring γ to δ (which presupposed $\Gamma \succ \Delta$) to preferring δ to γ (which implies $\Delta \succ \Gamma$). But the primary change concerns the belief about the decision situation itself, which makes the decision not to donate become OR. The decision-maker, thanks to awareness, is actually taking another type of OR decision, which does not anymore concern her friend smoking or not smoking, but her friend quitting or not quitting smoking. In this respect the end ω identifies a state of the world in which, as a consequence of the decision-maker's decision, the friend

quits smoking. This state of the world was out of consideration in the initial decision situation.

1.5 Responsibility and the intentionality of pro-social decisions

The enhancement of awareness can be interpreted as due to *reflection*, as the conscious or unconscious process of self-scrutiny of one's own values and ends. This process can be sustained both by deliberation (as a form of *self-reflection*) and discussion (as a form of *social reflection*). From the perspective of their reflective nature, deliberation and discussion have an analogous functioning and they both affect the decision-maker beliefs concerning the decision situation at stake. Deliberation and discussion condition the decision situation by setting the relevance of specific values and the salience of specific ends, so framing the interpretation of the OR or PS actions and consequences.

As far as PS decisions are concerned, reflection affects beliefs concerning the social features of the decision situation, by grounding the cognition of relevant social relationships and boundaries, which in turn define what the decision-makers may feel responsible for in performing the PS decision. Responsibility reveals a social aspect of awareness, in so far as the identification of relevant values and salient ends go together with the identification of 1) who is considered agent of social norms of cooperation; and 2) who is entitled to benefit from it. In other words, beliefs concerning relevant social boundaries determines the *subject of the social norm* both in terms of actions (who is entitled to interact for the benefit of the society as whole?) and consequences (*on who* consequences do matter?). The intentionality of the PS decision is shaped by these definitions to the extent that the mode and content referred to as the social good may change depending on the cognition of the significant social boundaries.

Discussion, thanks to its inter-subjective mode, plays a crucial role in the case of PS decisions. It sustains the development of a shared definition of the social community and of its values and ends. Let us recall the case presented in section 1.3 in order to describe

how the cognition of social contexts and relationships undermines the possibility that social heuristics specifies PS decisions. The execution of a given PS heuristics may depend on the beliefs concerning the boundaries identifying the group of cooperative agents and who is entitled to benefit from cooperation. In this respect, the evaluation of $\Gamma \rightarrow \gamma$ as PS (cooperative) depends on the fact that both the action and its consequences are interpreted as PS within the relevant social group. But let us make the hypothesis that $\Gamma \rightarrow \gamma$, while benefiting the group of the decision-maker's fellow members, is damaging (or at least not benefiting) a group of excluded agents. $\Gamma \rightarrow \gamma$ couldn't be anymore considered PS, in so far as, from the perspective of the society as a whole, that decision ends up generating more social exclusion. This situation is depicted in the following figure.

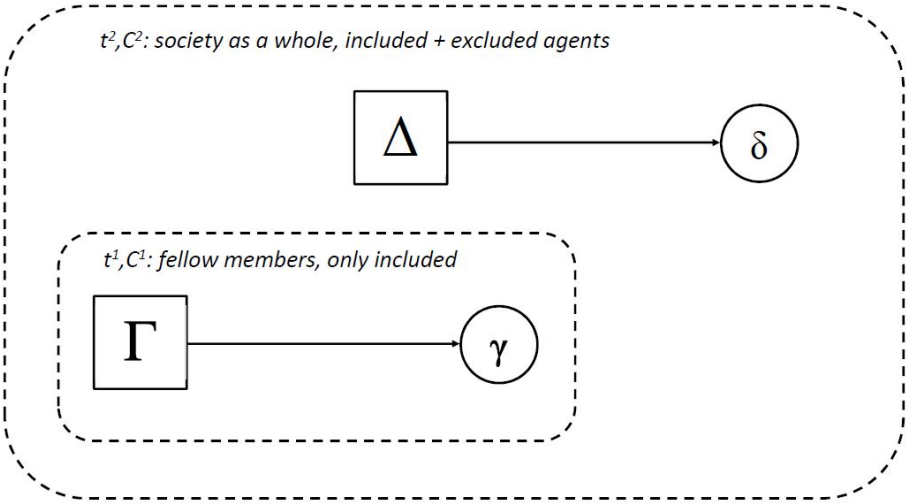

Figure 1.3: The emergence of responsibility

The decision-maker believes at time t^1 she is facing the decision context C^1 in which, for example, she can decide to cooperate with fellow members, doing the action Γ which has as a consequence the pro-social benefit γ to her fellow group. Possibly, in accordance to her social heuristics, she does so, on the ground of the belief she is deciding for the social good.

However, there is another option $\Delta \rightarrow \delta$ she is not aware of, which involves the

possibility to cooperate and share the benefit of cooperation also with agents who are not part of the original group of fellow members. This option depends on the assignment of a normative value to the alternative action Δ and/or on the recognition that consequences over outsiders δ and hence over the society as a whole matter and can be addressed as an end for PS decisions. Accordingly, the decision-maker may re-evaluate the original choice $\Gamma \rightarrow \gamma$ as anti-social, in so far as she can realize that keeping on not doing $\Delta \rightarrow \delta$ is detrimental to the society as a whole.

What could make the decision-maker pass from $\Gamma \rightarrow \gamma$ to $\Delta \rightarrow \delta$, if we exclude a change in (social) preference or an exogenous change in the dominant social norms? My hypothesis is that reflection, via deliberation and discussion (and in particular discussion with non fellow members) may change the beliefs of the (included) decision-maker so that the boundaries of the relevant social community, determining which actions and which consequences are PS for her, may change. In this way, possible cooperative interaction with non-fellow members and consequences over excluded, which was previously out of consideration, are attributed with normative values and ends so that they become available to decision-maker's cognition and practice.

An overall change in beliefs concerning the decision situation occurs. At time t^2 the decision-maker finds herself in the decision context C^2, for which the decision itself changes from one concerning the choice to cooperate (or not cooperate) with fellow members, to one between two alternative cooperative decisions: the one, pro-social only toward the group of fellows (and anti-social with respect to the society as a whole), the other pro-social with respect to the society as a whole. This process of awareness enhancement can be accounted for in terms of *responsibility*. The decision-maker, while acknowledging to be part of the society as a whole (and not only of its sub-group), becomes aware of possible (inter-)actions and consequences involving the others outside the group of fellows. Consequently, she gets responsible towards them with her decisions.

Thus, prosociality presupposes responsibility. Once the criteria defining which are the relevant social interactions and the recipients of the consequences of decisions are recognized, the normative definition of what is PS changes, from one ruling only over the

decision to cooperate with and benefit fellow members, to another in which cooperating only with fellow members is judged antisocial because it has detrimental consequences for the society as a whole. Responsibility triggers prosociality to the extent that the process of attribution of normative value and ends to actions and consequences involving agents out of the boundaries of the original peers' group entails the recognition normative "rights" to them together with their admission to the "cognitively relevant" social group.

The passage from t^1C^1 to t^2C^2 implies the passage from social preferences and norms, to *moral* preferences and norms. Their "universal" significance are conceptualized as superior as criteria for a PS decision, and identify $\Delta \rightarrow \delta$ as the PS option between the two. Since the reflective process determining the change in the criteria of identification of the relevant social group is an epistemic one, we can consider that, when it occurs, it necessarily implies that decisions adapt to the new normative conceptualization of the decision situation.

1.6 Conclusions

Deliberation and discussion, awareness and responsibility affect OR and PS decisions by framing their epistemic conditions. The analysis of these conditions sheds a light on aspects of OR and PS decisions, which cannot simply be reduced neither to un-reflected social heuristics, nor, in general, to given preferences for OR or PS options.

The current emphasis in moral psychology on intuitions and heuristics may lead to underestimate the epistemic function of moral deliberation in framing OR and PS decisions in actual normative contexts. By investigating the intentionality of OR and PS decisions, it gets possible to draw the limits of applicability of theories of cooperative and norm-compliant behaviours based only on intuitive heuristics. These limits coincide with context-dependency conditions and they are determined by beliefs concerning criteria of identifications of relevant options – both with respect to actions and consequences – which contribute to the definition of what is OR (or not) and what is PS (or not) in the decision situation. Unless these definitions are given and shared, any social heuristics can

apply. Deliberation and discussion can be viewed as reflective activities determining the emergence normative definitions *during* decision processes. Describing them as processes allows for identifying awareness and responsibility as necessary conditions of OR and PS decisions, while highlighting their normative essence.

To conclude, by taking into consideration intentionality of OR and PS decisions it has been possible to highlight an *ex-ante* framing role for moral reflection, deliberation and discussion. This in turn put us in the condition to show how OR and PS decisions necessarily involve *processes* of formation which requires time and interaction (also interaction with one-selves, i.e. deliberation) and which set the normative values and ends, which make relevant actions and salient consequences identifiable. These processes can also change the epistemic conditions in which the decision is taken, so that the decision faced in the beginning is not the decision answered at the end of the process (and *consequently* social preferences may adjust). Moreover, these processes – especially when discussion is concerned – may reshape beliefs concerning the relationships among persons referred in the decision and this may lead to the emergence of new OR and PS motives (and *consequently* normative references may change). New norms may emerge at the end of these processes, thanks to the recognition or redefinition of what is shared in the way in which members of the community understand and evaluates circumstances, contributing to their collective life.

CHAPTER 2

Normativity, commitment, and interactive intentionality

2.1 Introduction

The debate concerning the possibility to reconcile the "rule-based" and the "equilibrium-based" accounts of institutions (Hodgson, 2006; Aoki, 2007; Greif, 2006; Greif and Kingston, 2011) has been recently reactivated by Hindriks and Guala who tried to "unify" social ontology by explaining institutions in terms of "rules in equilibrium" (Guala and Hindriks, 2014, 2015). However, the replies to this attempt reconfirmed the distance between the two approaches (Binmore, 2015; Sugden, 2015; Hodgson, 2015). In particular, John Searle reaffirmed the irreducibility of "constitutive rules" as the basis of social ontology (Searle, 2015). This paper suggests an alternative strategy towards unification, by reflecting on Searle's epistemology of normative commitment.

The issue of unification derives from the circumstance that Searle's idea that institutions consists in the collective assignment and recognition of constitutive rules does not satisfy rationality criteria grounding the explanation of rule-following behaviours and the evolution of social norms provided by social sciences, and in particular by economics. Searle (1995, 2005, 2010) makes reference to the epistemology of collective intentionality in order to account for *commitment* (see also Gilbert, 1989, 2013). But this approach

does not satisfy the requirement of self-enforcement which accounts for rule-following in convention and social norms literature (Lewis, 1969; Schotter, 1981; Sugden, 1986; Young, 1998)

However, it must be observed that Searle is not engaged in providing an *explanation* in the social-scientific sense. The purpose of social ontology limits his account to the *description* of the necessary conditions under which institutional entities and social reality exist as collectively adhered. These conditions are summarized by the formula of "constitutive rules": provided that a group collectively recognizes (or "declares") that "X counts as Y in C", we have an institutional entity and hence social ontology (Searle, 2005, pp.9-10). This is not the explanation of institutions that social scientists are used to look for exactly because this is not meant to be such an explanation. As a description of conditions of social reality, Searle's social ontology does not account for the processes which makes those conditions exist.

Hindriks and Guala's unification of rule-based and equilibrium-based approaches to institutions relies on the "transformation" of "constitutive rules" into"regulative rules", i.e. on the possibility to translate the former in the language of the latter without semantic loss (Guala and Hindriks, 2014; Hindriks, 2009). Searle's definition of the two notions would not leave room to this strategy toward unification: on the one hand, *regulative rules* "regulate activities which can exist independently of the rule"; on the other hand, *constitutive rules* "not only regulate, but rather constitute the very behaviour they regulate, because acting in accordance with a sufficient number of the rules is constitutive of the behaviour in question" (Searle, 2005, p. 9). Ruling out this difference not only implies denying that status functions "create" social ontology, but it also leads to smooth over the epistemic dimension of normativity on the behavioural one.

The conception of institutions as equilibria in regulative rules is based on the solution concept of "correlated equilibrium" (Gintis, 2007, 2009) and on the idea that "players must be able to represent the equilibrium in symbolic form" in order to get to coordination in the context of games with multiple, asymmetric equilibria (Guala and Hindriks, 2014, pp.6-9). This idea is borrowed from Aoki (Aoki, 2007, 2011) who treats rules as

"coordination devices", i.e. publicly available representations of equilibria that induce *shared beliefs* concerning how the societal games have to be played. According to this view, rules correlate behavioural strategies only because they can condition beliefs by "specifying patterns of expected behaviours" (Greif and Kingston, 2011, pp. 25-28).

However, Guala and Hindriks disregard Aoki's distinction between the "behavioural dimension", in which the strategic choices are made, and the "cognitive dimension", in which beliefs concerning players' expectations about others' actions and expectations are formed. According to Aoki, institutions are an evolutionary process which connects the behavioural and the cognitive dimensions in a dynamic loop: on one hand, equilibria in individual behaviours are represented in rules which act as cognitive-media, fostering the formation of shared beliefs about the game; on the other hand, shared beliefs sustain equilibria in behaviours and hence sustain the behavioural regularities which identify the rule (Aoki, 2011, p. 23). The cognitive dimension is crucial in so far as it is the dimension in which mutual beliefs are formed as "commonly cognized salient patterns of the ways in which societal games are recursively played and expected to be played" (Aoki, 2011, p. 23). By dealing with institutions in terms of equilibria in regulative rules, Hindriks and Guala downplay rule's epistemic dimension and equate institutions to behavioural regularities.

I suggest an alternative strategy towards unification of social-ontological and social-scientific approaches to institutions. It is an alternative strategy to the extent that it does not imply the necessity to reduce constitutive rules to behavioural regulative rules. On the contrary, it is based on the idea that constitutive rules form shared normative beliefs which identify relevant actions and salient outcomes in institutional contexts, so that decisions and behaviours are consequently influenced. I develop this strategy from a reconsideration of Searle's conception of normative commitment. I rise two issues about it: 1) Searle's conception of normative commitment cannot account for *rule-following* decisions; 2) Searle's conception of normative commitment cannot account for constitutive rule's *formation*. The objective of unification is hence pursued by trying to integrate Searle's social ontology with an epistemology of rule-following capable to

account for constitutive rule's formation.

Searlian conception of normative commitment is grounded in the epistemology of *collective recognition* and *assignment* of status function, which in turn is based on collective intentionality. I consider that this epistemology cannot count as an epistemology of rule-following in so far as it can only refer to *unique* and *given* status functions [1]. The point is that the epistemic criteria that identify a status function as *that* specific status function must already be given in order to allow for its recognition and assignment to institutional objects, persons, or states of affairs. This is not a problem for social ontology *per se* – in so far as it is interested only in what is uniquely given (i.e. existing) – but it makes Searle's conception of normative commitment unable to account for rule-following. As it will be clarified, rule-following presupposes the cognitive availability of alternatives. By referring to *collective intentionality* as the epistemic basis of institutional reality, Searle rules out the possibility that an institutional entity can be understood and practiced in various ways across individuals. Consequently, the circumstance that institutions are subject to change due to non-compliant rule-following decisions and to processes of reconsideration of their deontological foundations remains out of consideration. Unless variety among the interpretations of status functions is admitted, we cannot account for constitutive rule formation through change.

I discuss two interconnected problems of Searle's conception of normative commitment. The first problem concerns Searle's notion of "function" and it is due to the fact that collective intentionality as the sole epistemic basis of commitment ends up presupposing commitment itself. I deal with this issue in terms of a "paradox of commitment". The second concerns the deontological character that Searle attributes to the notion of "status" and it is due to the fact that its *specific* deontological content cannot be justified by deontology alone. I deal with this second issue in terms of "incompleteness of deontology". These two problems limits Searle's conception of normative commitment , making it unable to deal with rule-following and constitutive rules' formation. To count

[1] The same can be said of "status function declarations" which Searle dealt with in a more recent work (Searle, 2010)

as valid representation of institutions from the social-scientific perspective, Searle's constitutive rule approach needs to overcome these limits and to illustrate how a common understanding and practice of constitutive rules can emerge and self-sustain, despite the multiplicity of individual interests and normative orientations within societies.

To account for rule-following and constitutive rule formation, it is necessary to allow for variety and change in status functions and to analyse under which epistemic conditions status functions are identified and selected among alternatives. To this purpose, I introduce a decision process – which I name *interactive intentionality* – through which individuals within institutions compare and contrast, via deliberation and discussion, alternative possible statuses' and functions' attributions to the same institutional entity. The main hypothesis is that interactive intentionality develops practical reason in an interactive way, so that individuals involved in it build shared normative interpretations by attributing values and ends respectively to actions and consequences.

This approach is based on trends in moral philosophy that reconsider deontological ethics as the result of historical, linguistic, and interactive processes (Habermas, 1990; Sandel, 1982). It hence supports the reintroduction of practical reasoning as a fundamental component of institutional life and evolution (Crespo, 2007, 2016; Velleman, 2009). I refer to interactive practical reason to account for endogenous processes of emergence of shared epistemic criteria which make specific status functions identifiable before they become objects of recognition and commitment. To illustrate how interactive intentionality can ground a common understanding and practice of constitutive rules, I develop a simple formalization. The hypothesis of interactive intentionality is aimed at representing the cognitive conditions of possibility of rule-following decisions and institutions' formation. It at the same time describes beliefs shifts and change in expectations concerning both others' norm compliant behaviours and others' expectations. The description of the effects of interactive intentionality on normative beliefs allows for dealing with the adherence to constitutive normativity also in terms of self-enforcement, so providing an epistemological ground for unification of social ontology and social scientific account of institutions.

The limits of Searle's conception of normative commitment are addressed in section 2.2 and 2.3, which respectively discuss "the paradox of commitment" and "the incompleteness of deontology". In section 2.4, I introduce multiplicity of values and ends and I explain why it is necessary to assume the possibility of conflict among them as a necessary condition of rule-following decisions. In section 2.5, I present the hypothesis of *interactive intentionality* as an interactive mode of practical reasoning and I show how it can be considered as the epistemic basis of institutional formation. Section 2.6 sums up and concludes by suggesting how interactive intentionality can be referred to obtain unification through a conception of normative commitment based on the idea that it depends on a prior inter-subjective acknowledgment of shared normative frames.

2.2 The paradox of commitment

Three primitive notions are required by Searle's social ontology to account for the subsistence of institutional entities: "collective intentionality", "function", and "status". Searle (2005) connects them pairwise. Collective intentionality intervenes in the collective recognition of "functions on objects where the object does not have the function, so to speak, intrinsically but only in virtue of the assignment of function." The assignment of the function goes together with the recognition of a certain status: "the object or person to whom the function is assigned cannot perform the function just in virtue of its physical structure, but rather can perform the function only in virtue of the fact that there is a collective assignment of a certain *status*, and the object or person performs its function only in virtue of collective acceptance by the community that the object or person has the requisite status"(Searle, 2005, pp. 6-8). I firstly discuss the connection between the notions of "collective intentionality" and "function". The connection between "function" and "status" is discussed in the next section (see below section 2.3)

The conceptions of "collective assignment of functions" and of "collective recognition" relate Searle's application of collective intentionality with the idea that constitutive rules specify institutional functions. Collective recognition is the basis of social ontology

in so far as the attribution of a certain function to a specific object depends on the fact that individuals forms collective intentions and adhere, in cognition and practice, to a common re-cognition of that object as having a certain function. Searle makes reference to the epistemology of collective intentionality in order to interpret collective recognition of institutional function. While assigning the function to the entity individuals join in collective intentionality, interpreting that object as having that function.

Assigning and recognizing a function is necessarily identifying what *purpose* is served by the object, person, state of affair. As clarified by Searle, the identification of any kind of function always presupposes the specification of a purpose:

> "[...] *functions are always intentionality-relative* [...] When we discover functions in nature, what we are doing is discovering how certain causes operate do serve a certain purposes, where the notion of purpose is not intrinsic to mind-independent nature, but is relative to our sets of values. [...] Where do the values come from? The clue that there is a normative component to the notion of function is that once we have described something in terms of function we can introduce a normative vocabulary. [...] To put the point succinctly, if perhaps too crudely, *a function is a cause that serves a purpose.* And purposes have to come from somewhere; in this case they come from human beings. In this sense, functions are intentionality-relative and therefore mind dependent." (Searle, 2010, p.59)

Hence, we can consider that the relation between Searle's application of collective intentionality and his notion of institutional function is mediated by the notion of "purpose". Any institutional reality possesses a status function in so far as it corresponds to a certain social purpose. The purpose is a necessary condition for the identifiability of the institutional function. It stands for the content of collective intentionality in the collective recognition/assignment of a status function.

In Searle's view the collective attribution and recognition of status functions comes with commitment. It is the linguistic nature of collective intentionality – and in particular of the speech act of "*status-function declaration*" (Searle, 2010, pp. 93 ff) – which intrinsically implies that people involved in collectively assigning the function are for

the same reason committed to accept and recognize it. Searle's application of collective intentionality is compliant with Gilbert's theory of "joint commitment" for which adhering to a common purpose in collective intentionality is at the same time being normatively committed to the purpose that is stated as a joint objective (Gilbert, 1989, 2013). It is the linguistic nature of the collective statement of a purpose at making, in Searle's view, collective attribution and recognition of institutional functions not reducible to cooperation in the sense of I-intentionality plus mutual beliefs (Searle, 2010, pp. 45-58)[2].

Hence, collective intentionality represents the epistemological basis of normativity in Searle's social ontology (Zaibert, 2003; Gilbert, 2007). Commitment to specific constitutive rules intrinsically derives from collective recognition of the function attributed to the institutional object or person. However, this conception of normative commitment does not account for *why* people actually converge on a specific institutional function. Such an explanation would depend on the possibility to identify alternative institutional functions and to represent individuals' choice among them. This possibility is out of the scope of the epistemology of collective intentionality, which can account for commitment in collective agency exactly because its intentional content in terms of purposes is already given and shared across individuals.

As observed in the introduction, being interested in social ontology, Searle is not engaged in providing an explanation of why people adhere to specific constitutive rules. However, the lack of this explanation exposes Searle's conception of normative commitment to the circularity of treating normative commitment to constitutive rules at the same as the cognitive cause and effect of collective recognition and assignment of status functions. In other words, lacking a representation of rule-following decisions – i.e. the decision whether to comply or not comply to the given constitutive rule among alternatives – Searlian conception of commitment entails a paradox in so far as it ends up presupposing commitment.

[2]See Tuomela and Miller (1988); Tuomela (1991, 2005). For a discussion of the alternative theories of collective intentionality in relation to conceptions of the individual agent see Davis (2003) pp. 130-149.

The collective recognition of a status function depends on the attribution of a function to the relevant object or person. The identification of that function depends on the circumstance that individuals cognize the purpose that define the function as the content of collective intentionality. In order to be an object of recognition and of declaration the function must be identified as serving a certain purpose (and not others). In other words, the purpose must have been already selected in order to be adhered in collective intentionality. Only if the purpose is already selected the relative function can be collectively recognized or declared as an institutional function. The function as such can be specified and referred to in collective recognition if and only if its intentional content is already cognized as one specific purpose that individuals share. But this is equivalent to assume that individuals are already committed to that purpose.

If the collective purpose is taken as given and shared, it means that any other alternative purpose has already been ruled out. Ruling out any other alternative (and potentially conflicting) purpose is an act which depends on commitment itself. This act of commitment is a precondition to the collective adhesion to the given institutional function and its subsumption under collective recognition and attribution. In this sense, Searle's conception of normative commitment ends up in a paradox: commitment, as a consequence of collective recognition and assignment of a function, presupposes the commitment towards the purpose which makes the identification of it as a function possible.

Searle needs to assume the selection of the relevant purpose behind the identifiability of the function as given in order to take collective intentionality as the epistemological basis of social ontology. But this epistemology results in a circular argument concerning normative commitment. In order to complement Searle's social ontology with a coherent theory of normative commitment, we have to account for how the common purpose behind the identifiability of the function is formed. This in turn requires admitting that institutional purposes can be characterized by multiplicity and re-conceptualising normative commitment in terms of rule-following decisions. Rule-following decisions presuppose the availability of alternatives. They are necessarily cognized as the choice

between adhering to a given norm or deviating from it complying to another norm. As far as constitutive rules are concerned, rule following decisions refers to the process of selection of a specific institutional purpose among alternatives.

2.3 The incompleteness of deontology

The paradox of commitment is embedded in a more general problem of *incompleteness* which concerns the deontological essence of institutional functions. The above mentioned connection between the primitive notions of "function" and "status" derives from the consideration that recognizing and assigning and institutional function is necessarily recognizing and assigning a normative "status". In Searle's words, the institutional object or person "cannot perform the function in virtue of their physical structure alone, but only in virtue of the collective assignment or acceptance of the object or person as having a certain *status* and with that status a function" (Searle, 2005, pp.7-8). The status is specified in terms of *deontic powers*, i.e. the formal definition of duties and rights, providing people within institutions with *desires-independent reasons for action*:

> "The essential role of human institutions and the purpose of having institutions is not to constrain people as such, but, rather, to create new sorts of power relationships. Human institutions are, above all, *enabling*, because they create power, but it is a special kind of power. It is the power that is marked by such terms as: rights, duties, obligations, authorizations, permissions, empowerments, requirements, and certification. I call all of these *deontic powers*. [...] They are structures of power relationships. [...] how exactly do these power relations function? The answer, which again is essential to understanding society, is that institutional structures create desire-independent reasons for action. To recognize something as a duty, an obligation, or a requirement is already to recognize that you have a reason for doing it which is independent of your inclinations at the moment. [...] the creation of the general field of desire-based reasons for action presupposes the acceptance of

a system of desire-independent reasons for action." (Searle, 2005, 10-11)

The notion of "deontic power" is inherent to the notion of "status" in Searle's definition of status function: an institutional entity is identified as possessing a certain status in so far as it displays certain deontic powers; but also, the other way round, certain deontic powers identifies that person or object as having a certain status. The function of the institutional entity (and hence its implicit purpose) reflects its status and relative deontic powers. By specifying "desire-independent reasons for action", statuses and deontic powers define a set of *possible* actions which are attributed with normative value: duties and rights assigned to the institutional entity (or connected to its use) as defined in terms of potentiality of action and linked to the functional definition of its purpose.

However, no explanation is provided concerning why the power implied by the status is that normative power – enabling certain actions – and not another – enabling other actions (Sánchez-Cuenca, 2007; Viskovatoff, 2003; Zaibert and Smith, 2007). Being interested in the analysis of formal and essential (irrespective of differences) properties of the normative structures of institutions, Searle's ontological description does not need to specify why determined deontic powers and not others are attributed to (some) persons (and not others) or objects, in the moment of the collective recognition of the institutional status function. But in this way, Searle's deontology lacks the possibility to account for its own *formation*. In this sense, I consider that Searle's deontological theory of institutional normativity is incomplete: there is no way to justify, within the scope of deontology alone, why *that* status attribution concerns *that* power and not another, nor to account for the emergence of *that* specific power against possible others.

The issue is analogous to the one concerning commitment, which, as we have seen above, presupposes the previous identification of a given purpose (and hence the process of selection for taking it as given). Here the possibility to recognize the status depends on the fact that the powers – the set of potential actions attributed to the entity – are already defined and shared. Social ontology assumes purposes and powers as given in order to ground the cognitive identifiability of status functions to which collective recognition applies. The attribution of status function depends on the collective recognition and

assignment, but this collective recognition can attach to status function as its object only because the status and related function are made identifiable as the bearer of specific powers and purposes. The epistemology of collective intentionality presupposed by Searle needs to take the normative definitions of powers and purposes as given, but there is a deeper epistemological level in which they are cognized as multiple alternatives which are compared, contrasted, discussed, selected. I consider the epistemology of these processes of rational scrutiny and decision concerning normative powers and purposes is the epistemology of rule following and norm formation.

The opportunity to compensate social ontology with an epistemology capable to account for rule-following decisions and norm formation can be highlighted by taking into consideration the case of *intentional deviation* from constitutive rules. This case cannot be accounted within Searle's conception of normativity and of normative commitment. From the social ontological point of view, as far as a constitutive rule is defined, it is collectively adhered. On the contrary, intentional deviation from constitutive rules calls for admitting the possibility that individual treats as a possible object of choice alternative purposes and status attributions to the same institutional entity. However, it must be observed that intentional deviation from rule presupposes the acceptance and recognition of the given constitutive rule and of the institution as such. If the given status function was not recognized by deviant individual, it could not be addressed as the object of the envisaged change through deviation. Individuals who deviate from a given constitutive rule must recognize it in order to think at and pursue its possible change.

Moreover, intentional deviation from constitutive rule must be treated as a rule-following behavior. It indeed presupposes the commitment to an alternative normative definition of the same institution. It entails the alternative attribution of normative power and purpose to the same institution and the alternative interpretation and definition of the status function. However, in order to count as an alternative definition of a status function it has to be object of collective recognition. Indeed, individual deviation necessarily presupposes that the deviant agent thinks at herself as part of a collective

which assign the alternative interpretation of the status function to the given institution.

Intentional deviation from constitutive rule highlights a type of "negative commitment" which cannot be dealt with within Searlian conception of commitment. The epistemology of collective recognition is only "positive" since it is confined to defined status functions. It presupposes that the selection of purposes and powers which make that status function identifiable has already been brought to completion. But this excludes the possibility that institutional agents could make reference, in cognition and practice, to alternative normative definitions of those purposes and powers, and deviate from given constitutive rules. As a consequence, Searlian treatment of deontological powers and purposes hides a peculiar aspect of asymmetry concerning rule-following and normativity in general. This asymmetry has two aspects. First, only admitting negative commitment we can account for normative commitment in general. Second, only admitting the possibility of deviating from given normative assets we can account for norms formation through constitutive rule change.

As far as the first aspect is concerned, unless we allow for the possibility of intentional deviation from rule, we cannot account for norms' recognition and commitment itself. From a logical-epistemological point of view there cannot be normativity without alternatives. The decision to rule-follow is necessarily the decision not to deviate from it. If no alternatives would be available in cognition, that behaviour could not be considered as the decision to follow a norm, because it would be determined only by the availability of only one, unique, and given option. Only admitting the availability of alternative normative references, we can account for rule following and hence for the adhesion to norms in institutions. In order to represent constitutive rule-following, and to overcome Searlian paradox of commitment, we have to account for individual decision among a plurality of institutional purposes and powers.

As far as the second aspect is concerned, only admitting that within the society there are multiple possible alternative interpretations of status functions in terms of purposes and powers, we can account for processes which bring institutions into existence. Institutions emerges as the result of processes of formation which presuppose deviation

from norm as a means to institutional change and the composition of shared definition of status functions via conflict resolution. Incompleteness of Searle's deontology can be compensated by allowing for these processes of formation .

In the rest of the paper I suggest an epistemology which is alternative to collective intentionality and which makes a description of the epistemic conditions behind both rule-following decisions and of constitutive rule formation available. This epistemology needs to account for processes and interactions through which institutional powers and purposes are cognized and selected among alternatives. It suggests a way to make social ontological account coherent with social science's explanations of institutions as self-enforcing equilibria, by attributing a framing role to these processes of selection and decision concerning the attribution of normative value to actions and institutional purposes.

2.4 Variety and conflict in status functions

In order to complement Searle's social ontology with an epistemology capable to describe conditions of rule-following and norms formation, I allow for multiplicity of institutional powers and purposes. This entails that alternative status and functions, depending on the definition of purposes and powers behind them, can be attributed to the same institution in a given moment. Within a given social context, we can have that: a) group A considers that X and hence Y serve a certain purpose, while group B attributes to X and hence to Y an alternative purpose; and/or b) group A considers that X and hence Y have a certain power, while group B attributes to X and hence to Y an alternative power. In both cases, the two groups collectively recognize that "X counts as Y", but for group A Y is Y^1, while for group B Y is Y^2. Note that, from an ontological point of view, the status function remains identical to itself. It only changes the normative commitment of the two groups and it reflects the respective interpretation of purposes and powers.

The two described conditions represent circumstances of potential *conflict*. Such

a conflict can be interpreted both as a decision-making problem or as social cohesion problem. The first problem is linked to the above mentioned issue of "intentional deviation from rule". It is the rule-following problem consisting in the individual's conflict concerning whether a) to accept and get committed to a given institution, despite her normative commitment toward an alternative interpretation of the status function or b) to deviate from it, complying to the normative definition of powers and purposes recognized within her social group to the purpose of constitutive rule change. The second problem, the social cohesion issue, derives from social conflict: how can different groups may get to recognize a common shared interpretation of the same institutional status function? The issue here concerns how do institutions may adjust for differences across social groups so that the status functions are representative of shared interpretations of institutional powers and purposes. The two issues shed a light on the problem of institutional *formation*.

I deal with these circumstances of conflict by analysing the epistemic conditions behind them. With the notion of "epistemic conditions" I refer to beliefs concerning allowed actions and licit consequences in the decision situations. We can consider that Searle's deontic powers and institutional purposes respectively refer to the normative determination of the two sets of allowed actions, on the one hand, and licit consequences, on the other. Powers, defined in terms of potentiality of action, attribute intrinsic deontological value to actions. Purposes envisage as expected outcomes what the institution is aimed at. On the ground of this interpretation of Searlian cognitive preconditions of status function, I consider that individuals involved in the above mentioned circumstances of conflict engage themselves in a deliberative process concerning the attribution of one among multiple, alternative intrinsic values to relevant actions and the interpretation of consequences of decisions in terms of one among multiple, alternative ultimate ends. This deliberative process is introduced to describe how the cognition of shared powers and purposes may emerge from conflict. The process of formation of shared interpretations of powers and purposes is viewed as dependent on "rational" choice about values and ends attributions to actions and outcomes.

I make reference to philosophical discussion of "incommensurability of values" to characterize the relation between values and ends (and the relations within each of the two categories). In particular, the deliberative decision model that follows rests on the basic distinction provided by Chang (1997, pp. 4-7) between values – as the systems of valuation which may happen to be commensurable or not – and the bearers of value – the objects or state of affairs which, according to that system of valuation, may happen to be comparable or not. I name the second "ends" to suggest the idea that their cognition is the result of an act of projection towards the world. The values-ends distinction allows for a characterization in terms of intentionality of the normative attribution of values and ends to actions and consequences. The attribution of intrinsic value to an action entails the interpretation of its envisaged consequences as correspondent to an intrinsic end. In this sense, the value defines a *mode* of normative reference which is projected on the objects or state of affairs as the concrete outcome of the actions. The end represents the intentional *content* of the value. This intentionality interpretation of the relationship between values and ends parallels the relationships between the notions of "status" and "function" (and between "powers" and "purposes") in Searle's treatment of constitutive rules.

The deliberation concerning values and ends describes the epistemic context in which the status functions is selected as a specific status function through the definition of powers as possibility of actions and functions as envisaged purposes. Thus, the introduction of values and ends is aimed at making explicit the processes of interpretation and definition of institutional powers and purposes. Specifically, the conflict among alternative powers attributions and among function attribution can be dealt with in terms of conflict among individuals interpretation of actions in terms of intrinsic values and consequences in terms of ends. Therefore, the resolution of such a conflict represents the way in which individuals get to a shared definition of purposes and powers, which paves the way to the collective recognition of the status function.

Conflict can concern both values and ends affecting the possibility that individuals share a common interpretation of the constitutive rule, and hence adhere to institutions

in decisions and behaviours.

When conflict concerns values, individuals believe that the actions to which they respectively attribute intrinsic normative value are mutually non compatible. On the other hand, when two values are compatible individuals believe that the actions to which they respectively attribute intrinsic normative value are compatible. When two values are compatible, there exists a common meta-value which is compatible with each individual value: individuals believe that the actions to which they respectively attribute intrinsic normative value can be jointly realized.

When conflict concerns ends, individuals believe that the envisaged state of the world that they respectively put as a purpose of collective action are mutually non compatible. On the other hand, two ends are compatible when individuals believe that the envisaged state of the world that they respectively put as a purpose of collective action are compatible. When two ends are compatible, there exist a common meta-end which is compatible with each individuals' ends: individuals believe that there exists a state of the world in which the two individually envisaged states of the world can be jointly obtained.

Values and ends are linked by a double relation. Each value implies an end: a state of the world which come to existence after the action to which is attributed value is realized. Each end presupposes a value: an action which is attributed with a value because its envisaged consequence is viewed as a purpose. Given this relationship, when two values are compatible the respective ends are compatible and *vice versa* when two ends are compatible the respective values are compatible. On the other hand, when two values are not compatible the respective ends are not compatible and *vice versa* when two ends are not compatible the respective values are not compatible[3].

To schematise, let us introduce the following notation and graphics. Two values Γ and Δ are compatible (Γ C Δ) when individuals believe that the actions to which they respectively attribute intrinsic value can be jointly realized. In that case, there exist a

[3]The relation of compatibility (and not compatibility) as involved in practical reasoning can be interpreted in terms of commensurability of values and comparability of ends. See Millgram, Regan, and Raz in Chang (1997)

common meta-value Ω compatible with individuals' values (Ω c $\Gamma \wedge \Omega$ c $\Delta \Rightarrow \Gamma$ c Δ) and individuals believe that the actions to which they respectively attribute intrinsic normative value can be jointly realized. Vice versa, they are non-commensurable when individuals don't believe so (Γ -C Δ).

Two ends γ and δ are compatible (γ c δ) when individuals believe that the envisaged outcomes of decisions can be jointly realized. In that case, there exist a common meta-end ω compatible with individuals' ends (ω c $\gamma \wedge \omega$ c $\delta \Rightarrow \gamma$ c δ) and individuals believe that there exists a state of the world in which the two individually envisaged states of the world can be jointly obtained. Vice versa, they are non-compatible when individuals don't believe so (γ -c δ).

Conflict is characterized by the epistemic circumstance in which individuals believe that their respective values and ends are mutually non compatible. This condition, for which γ -c $\delta \Leftrightarrow \Gamma$ -C Δ, is depicted in Figure 2.1.

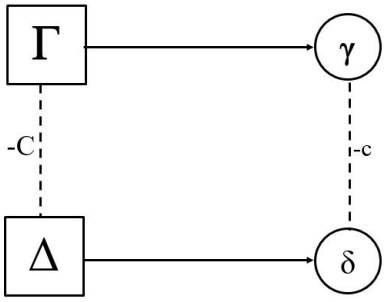

Figure 2.1: Conflict as non compatibility of values and of ends

2.5 Institutional formation through *interactive intentionality*

Variety among values and ends – as among alternative interpretations of institutional statuses and functions – calls for deliberative decisions capable to set both individual and collective resolutions in the case of conflict. Such deliberative decisions are grounded in *practical reason* as a basic moral faculty through which rational scrutiny and deci-

sions concerning values and ends are developed. I follow Habermas' *discourse ethics* in attributing a linguistic and performative nature to practical reason (Habermas, 1990). Habermas reinterprets Kantian deontological philosophy, by grounding normativity – Kant's "categorical imperative" – in dialogical, interactive performance of practical reason which gets to identify the universally valid norm through rational discussions of individual normative claims. Likewise, I treat the process which leads individuals to select and share the definition of institutional powers and purposes as a rational discussion developed through the inter-subjective performance of practical reason. Consequently, I consider that once this discussion comes to success, thanks to the universal nature of the obtained solution, it is normatively and cognitively binding, grounding collective recognition of the status function.

The conditions of conflict described in the previous section (section 2.4) are representative of institutional contexts in which the normative content of status functions – both with reference to powers and to purposes - is subject to debate. I introduce the hypothesis of *interactive intentionality* which I define as an interactive mode of practical reasoning applied to the selection of values and ends among alternatives. The analysis of the epistemic conditions under which conflicts concerning normative values and committed ends are overcome shows how interactive intentionality may get to shared normative definitions of institutional powers and purposes, contributing to institutional formation. More precisely, interactive intentionality can be thought as a deliberative discussion process in which individuals access their respective values and ends, compare and contrast them inter-subjectively, and eventually get to a shared identification of which actions and which consequences of action are to be held as normatively valid.

The resolution of institutional conflict is viewed as depending on the circumstance that individuals, through interactions in practical reasoning, may get to know with normative values and ends which point at possible available actions and consequences, not considered in the initial epistemic condition. With reference to the conflict situation depicted in Figure 2.1, we may consider that, at time *t'* individual A and individual B mutually believe that their respective values and ends are not compatible. Specifically, A

believes that the state of the world that the end γ point at as consequences to be pursued and the set of normatively possible actions identified by value Γ are not compatible with those entailed by B's normative value Δ and end δ. B believes the same about A's normative assumptions. We are hence interested in describing under which epistemic conditions this conflict could be solved.

On the one hand, it may be the case that through deliberative discussion individuals start comparing their respective deontological attribution of values to actions. Through the comparison of values attributed to actions, many possible joint actions are put under scrutiny as object of possible attribution of normative (deontic) value. It is hence possible that thanks to deliberative discussion, besides many non compatible values, individuals become aware of a previously not considered value or that a new normative principle is identified as relevant on that occasion (legal innovation). In this way, it can be the case that a value Ω emerges and individuals realize that it is compatible with each of their respective original values, so that is embraced as "universal". Hence, there exist a joint action, compliant to the emergent normative principle, on which individuals can converge. The success of this deliberative process entails that at time t" individuals have changed their beliefs concerning the decision situation in such a way that now they view their respective values as compatible and subsumed under a common possible value Ω. The value Ω points at a joint action which satisfies both the originally conflicting value attributions. The process of conflict resolution through the emergence of a common compatible value is represented in figure 2.2.

On the other hand, it may be the case that through interaction individuals start investigating the practical options available to them. This process may entail that many states of the world, as outcomes of individuals' interactions, are taken in consideration and evaluated as possible collective normative ends. Through this discursive procedure it may happen that previously not considered outcomes become salient (states of the world which previously was out of the set of feasible outcomes for both individuals) or new combinations of feasible outcomes due to a creative process of knowledge sharing (technological innovation). Among these outcomes it can be the case that an end ω

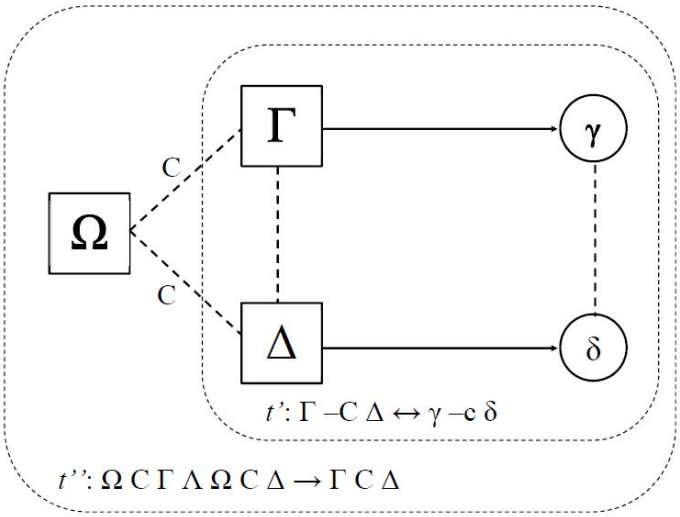

Figure 2.2: Conflict resolution via emergence of a compatible value

emerges and each individual realizes that it is compatible with their respective original ends. There exists a state of the world in which both original collective ends can be pursued at the same time. This means that at time *t"* individuals have changed their beliefs concerning the decision situation, in such a way that now they view their respective ends as compatible and subsumed under a common possible end ω. The end ω points at a state of the world in which both the original conflicting ends are realizable. The process of conflict resolution through the emergence of a shared compatible end is represented in figure 2.3.

The two modes of conflict resolution identify necessary epistemic conditions behind the adherence to a new collective attribution of normative relevance to actions and consequences. They hence illustrate an epistemology capable to account for the formation of shared interpretation of powers – via shared values – and purposes – via shared ends. Accordingly, interactive intentionality is viewed as a precondition of collective recognition and assignment of status function to object, persons and state of affairs in Searle social ontology. It moreover offers the opportunity to describe the epistemic conditions behind rule-following decisions as determinant of normative commitment to constitutive

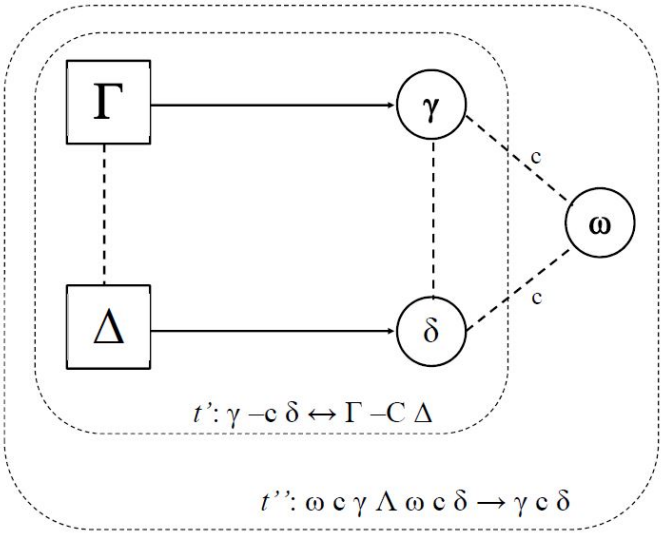

Figure 2.3: Conflict resolution via emergence of a compatible end

rules. Given the process and interactive nature of interactive intentionality, individuals commit themselves to the result of the process exactly because they engaged in it, putting interactively under discussion their own position. The "universal" nature of the obtained solution is such that once the individuals get to know it they are convinced by its rational superiority and committed to implement it.

Figure 2.4 illustrates how, once the shared normative interpretation of relevant actions and consequence has emerged, it entails a reconfiguration of the whole context of normative reference. This process can be interpreted as grounding the formation institutions,through conflict resolution and constitutive rule's change.

The fact that interactive intentionality entails beliefs' change about the normative valid actions and consequences can be referred to bridge the gap between Searlian social ontology and conceptions of institutions as behavioural equilibria based on the mutual expectation that others will comply to norms. Interactive intentionality can be interpreted as acting over frames, so that the relevance of certain actions and the salience of certain consequences change, following to the new normative interpretation of the decision situation. Individual's expectations concerning both the others' compliance to

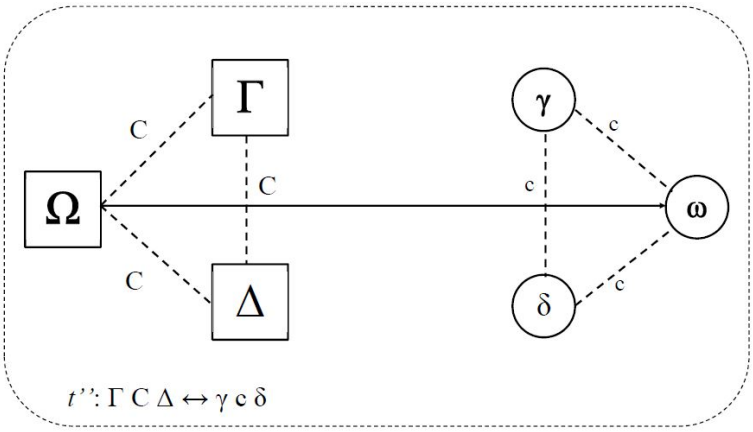

Figure 2.4: Formation and recognition of new purposes and powers

the constitutive rule and others' expectations about her own decision to comply to it, may accordingly be positively affected. It is hence more likely that, once interactive intentionality comes to an agreement and produces a shared normative interpretation of the decision situation, individuals stick to this normative interpretation and follow the rule they gave to themselves.

2.6 Conclusions

The interactive intentionality hypothesis accounts for institutional formation as a consequence of collective, inter-subjective elaboration of shared values and ends attributed to actions and consequences. The interactive process through which this elaboration takes place accounts for a change in individual's beliefs concerning compatibility of actions attributed with intrinsic value and states of the world envisaged as an end that ought to be pursued. In this way cognitive criteria which makes status and functions identifiable – collective purposes and deontological powers, respectively – are formed and constitute the basis of collective acceptance and recognition.

To the extent that interactive intentionality is implemented as an interactively process, it commits individuals to its result. Institutions are (also) formed through deliberative discussions which commit individuals who take part in it, because the normative

solution they get through interaction can claim to be "universal". Moreover, we can consider that, by changing the beliefs concerning relevant actions and salient consequences, interactive intentionality changes the individuals expectations concerning the others' compliance to the norm and the others' expectations concerning the decision-maker compliance to the norm, so allowing for an account of constitutive rule in terms of self-enforcement.

By discussing Searle's foundation of normative commitment into status function's collective recognition it has been possible to highlight epistemic interactive processes acting behind collective intentionality in institutions. I underlined that, as far as normative commitment is concerned, it must be clarified how specific powers and purposes are selected among alternatives. Without this clarification we remain within the limits of deontological ontology, for which it is not possible to account for how institutions emerges and exists as specific institutions. The specificity of institutions and the fact that they are intrinsically subject to evolution via conflict resolution and normative change are ontological traits which remain (legitimately) out of Searlian constitutive rule approach.

The attempt to overcome the incompleteness of deontology, by looking for epistemic conditions of constitutive rules' formation, opens the possibility to think at rule-following behaviours as grounded in practical reason modes of decision. Interactive practical reasoning constitutes an alternative to both commitment-based and enforcement-based conceptions of norms compliance and suggests a way to reconcile the two. Interactive intentionality, as a processual and inter-subjective performance of practical reason, acts behind both deontological and instrumental rationality, constituting a cognitive basis for both commitment and enforcement. It accounts for the formation of shared cognition of normative values and ends as respectively attributed to actions and states of the world viewed as their consequences. Interactive intentionality it is not only presupposed by norm compliant behaviours, but it also provides with shared epistemic references the definition of the "incentives" which constitutes the cognitive references of self-enforcement.

From the discussion of Searle's conception of normative commitment, it emerges a

conception of social ontology as intrinsically evolutionary. Status functions are what they are because they become what they are through rule-following decisions and constitutive rule's formation. We recognize and accept institutions also because we can believe that we can change them, because we believe that it is possible they can be different. This is part of our commitment, and it makes our rule-following a free act of decision, necessarily involving alternatives.

CHAPTER 3

The effect of ethics meetings on risk-taking behaviour: an experiment

In collaboration with Fracensco Feri (Royal Holloway University of London) and Caterina Giannetti (University of Pisa)

3.1 Introduction

In the aftermath of the financial crisis, three objectives emerged for financial market reforms: 1) enhancing financial stability and managing systemic risk (macro-prudential regulation); 2) rebuilding financial institutions (micro-prudential regulation); and 3) making the financial sector work for the benefit of financial users and society as a whole (Wehinger et al., 2012). While the first two objectives were the focus of regulators, the third objective has received comparatively little attention.

One reason for this can be identified in the circumstance that the issue exceeds the boundaries of traditional approaches to efficiency and regulation of the financial sector, involving social responsibility and businness ethics. However, it is now widely accepted that financial crisis was not just crash (Zingales, 2015), and failure by banks and bank agents to meet ethical values played a significant role in it. This acknowledgment has generated interest in the provision of culture and ethics within the financial sector, as a measure to contrast widespread tolerance of dishonest behaviours (Morris and Vines,

2014; Klooster and Meyer, 2016; Cohn et al., 2014, 2015). Despite culture and ethics are recognized as important determinants of human and organizational behaviour (Guiso et al., 2015), their fostering through policy is often viewed somewhat impractical (Wehinger et al., 2012). With this paper we contribute to this issue, by studying in the laboratory the conditions and the practices which favour the emergence of pro-social norms and behaviours in bank activities. In particular, by focusing on risk-taking decisions, we investigate whether individual moral deliberation – which we deal with in terms of *reflection* – and ethic discussion among peers – which we conceptualize as *ethic meetings* – are effective in reducing social costs of excessive risk taking by financial agents.

To this purpose, we propose a novel experimental setting which let us structure a decision-making situation in which there is monetary conflict of interest between the decision-maker and the passive receiver of the consequences of the decision. Our aim is to simulate *risk-taking for others* as representative of risk-taking decision in bank contexts. Through manipulation of decision times and discussion among peers, we assess the effect of deliberation and ethics meetings on individual risk-taking when the decision-maker has to choose between two risky gambles, one of which always implies a better outcome for the investor, but also exposes the client to higher risk. More precisely three treatments are applied: a "*Baseline*" treatment, in which subjects are asked to choose within a relatively short time; a "*Reflection*" treatment, aimed at assessing the effect of deliberation over the risk-taking decision at stake; a series of "*Ethics-meeting*" treatments (differing with respect to communication setting) aimed at assessing the effect of peer discussion concerning choice's intentions and consequences. The hypothesis we test is whether, in the context of risk-taking for others, reflection and ethics meetings make responsible choices more likely, minimizing risks for clients and for the society as a whole.

The motivation behind the research stems from the consideration of significant shortcomings of both incentives-based and rules-based attempts to promote ethical controls over bank activities. On the one hand, the permanence of a "risk culture" in banking

is actually reflected and reinforced by prevailing compensation practices, which tend to over-emphasize (short-term) revenue generation relative to downside risk. In this respect, standard incentive schemes in the financial sector appear ill-suited to resolve agency problems, and often encourage people to undertake investment strategies which reward opportunistic behaviours and create socially excessive risk-taking (Young et al., 2012; Awrey et al., 2013)[1]. Moreover, it has been shown that, when prevailing business norms and personal ethical commitments are not aligned, and advisors are incentivized to give bad advice, honest recommendations decrease (Gneezy, 2005; Burks and Krupka, 2012).

On the other hand, conventional normative approaches to financial law and regulation (e.g. bank capital ratios) appear to have limited impact as a means to directly address market failures (Awrey et al., 2013). In the last years, the financial services industry has produced several "code of conduct", "code of ethics" and "principle of best practices", but with few concrete results. A prominent example is the "Code of Ethics and Standards of Professional Conduct" of the Chartered Financial Analyst Institute. The main problem is that the business culture that these normative codes aim to promote is contrasted and overcome by other countervailing behavioural norms within the bank. For example, the "norm of self-interest" generates pressure to exploit counterparts interests, and is often confirmed by existing incentives schemes (Young et al., 2012). In addition, other tacit norms may lead to neglect the social impact of individual risk-taking behaviour and encourage managers to increase the volatility and, therefore, riskiness of the institution's asset profile (e.g. "too big to fail", see Awrey et al., 2013). In general, it has been emphasized how the subsistence of a risk-culture in the banking industry may depend on the transmission of negative values and norms among the members of the bank community, such that the diffusion of irresponsible behaviours is favoured (Lo, 2015).

[1]When incentives are devised as performance compensations, lending volume rewards lead to high underestimation of risks, high acceptance rates, and high default. By contrast, high powered incentives are more effective at generating the correct assessment of loans (Cole et al., 2015) but they lead the decision-maker to neglect the consequences on others if there are monetary conflicts of interest with the clients (Andersson et al., 2013).

The need for new control instruments, besides incentives and rules, justifies our attempt to investigate the effects of ethics meetings as an internal and self-sustaining way to enhance bank social responsibility. Our aim is to verify whether the promotion of ethic awareness can foster ethical decisions in risk-taking for others. One relevant feature of our decision context, which resembles many financial decisions, is that there is not strategic interaction (at least at the moment of the decision) between the decision-maker and the individual who is affected by the decision. Indeed, our decision-maker is free to choose, without any further constraint, among two alternative risk options. One is potentially more lucrative for himself as the investor, but also likely cause loss of earnings for the client. The other is safer for the client and potentially less lucrative for him. This feature makes risk taking for other essentially an ethical decision, for at least two basic reasons: 1) the bad consequences of risky decisions are entirely charged on other subjects, who are not involved in the decision itself (and who do not benefit from potential extra-gains); 2) there is an asymmetry between who takes the decision and who is affected by it, such that the passive receiver of the consequences of the decision is unable to influence it.

We develop our hypothesis on the ground of Dual Process Theory of moral judgment (Greene, 2014), according to which some moral judgments are driven by automatic responses, while others require an effortful processing involving the conscious evalutation of opportunities. By distinguishing the influence of the reflective decision process (i.e. longer-deliberation times and peer-discussion) from that of faster modes of decisions, we identify the effect that the promotion of moral reflection and ethics meetings has on the emergence of other-regarding norms. In particular, we expect that by (auto)framing deliberation and peers-discussion individuals will recognize the normative dimension of their decision, thereby correctly assessing its relevant consequences. The main idea is that, after decision makers get, via deliberation and discussion, the relevant knowledge concerning the decision problem and their position with respect to it, others payoffs become salient. Thus, decision-makers may realize their other-regarding dispositions and possibly experience guilt aversion.

The paper is organized as follows: section 2 positions our contribution in the context of experimental studies on the role of moral norms and values in the emergence of pro-social behaviours and social norms; in section 3 we present the experiment design; section 4 illustrates and discusses main results; section 5 concludes.

3.2 Literature review

In order to frame the issue of risk-taking behaviours in banking, we focus on decisions under risk made by a decision-maker who chooses for another individual, i.e. risking (also) other individual monetary payoff. In order to do this, we set up a new experiment which confront our subjects with a sort of risky dictator game (Bohnet et al., 2008). In particular, in our experiment the decision-maker has to decide between two risky gambles: a "safer" investment, which always produces a certain amount of money for the decision-maker, and a "riskier" investment, which potentially allows him to gain more money but also exposes another individual to the risk of getting a lower payoff. In other words, irrespective of the level of individual risk aversion, the decision-maker has an interest in making the riskier investment, while his partner has a preference for the safer investment. Moreover, from the point of view of the society as a whole, the safer investment would be preferable.

In this kind of situation, fairness theories (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) would predict a greater attraction of the safer option, when the riskier option implies unequal payoffs. Bolton et al. (2008) indeed show that individuals tend to be more resistant to imposing risks on others than on themselves. However, they also show that individuals tend to choose more frequently the riskier option when the safer option implies unfavorable inequality, suggesting that social comparisons can significantly affect risk taking.

The literature also singles out that when individuals are not accountable for their choices, they tend to be less risk averse when investing other's people money (Pollmann et al., 2014). Moreover, Andersson et al. (2014) show that individuals are less loss averse

when managing risks for other people. They explain this circumstance by referring to a more "rational" elaboration of the decision made on behalf of others, which reduces the affective bias associated to loss aversion. In general, economic rationality should lead the decision-maker to selfishly exploit this type of decision situation, against more responsible and other-regarding decisions.

However, the experimental literature studying the effects of social and moral norms on dictator-games suggests that norm-compliance explains a considerable amount of variation in other-regarding behaviour. For example, Krupka and Weber (2013) show that an indicator of social norms they derived from a game-based elicitation method accurately predicts behavioural change across variants of the dictator game (e.g. the "bully version" of the dictator game). In the same vein, Schram and Charness (2015) examine the effects of social norms on dictator choices by letting peers, who do not have any financial payoff at stake, provide normative advice to players. They find that when one's decision is public, there is a tendency to choose more in line with the received advice. The compliance to a moral norm is thus induced through the advice given to the dictator. In line with these studies, we make the hypothesis that social norms may help to explain other-regarding and pro-social behaviours. However, differently from them, we allow the relevant social norm to become salient through decision-makers' discussions, which should promote the comprehension of the risky feature of the decision at stake and its normative implications. We hypothesize that discussions may foster norm-compliance acting on two types of expectations, respectively defined by Bicchieri (2005) as *empirical* and *normative*. Empirical expectations concern how many people will adopt a determined behavior. Normative expectations concern what others think it is *ought to do* in the decision situation (see Bicchieri and Xiao, 2009).

More precisely, we contribute to this literature by showing that communication, which only intervenes among decision-makers, increases the comprehension of the risky feature of the decision at stake, so that they become aware of a social norm opposing risking on others. Accordingly, we make the hypothesis that ethic meetings, in the form of peer discussions, may reframe subjective beliefs concerning relevance boundaries, so

that the decision-maker may better evaluate his role in the decision situation and consider himself subject to a wider social norm including the point of view of subjects external to the peer groups. Thus, the decision-maker may start feeling responsible for consequences of his decisions previously overlooked or under-evaluated, because "external" to the cognitive frame which identifies relevant consequences in the initial decision context.

As highlighted in the introduction, this hypothesis is grounded on the Dual Process Theory of moral judgment (Greene, 2014). In particular, our experimental setting is analogous to the right-wrong decision-context analysed by Gunia et al. (2012), who account for the connection between ethical choices, on the one hand, and deliberation and discussion, on the other. Right-wrong decision is a specific type of moral decision between an intrinsically ethical course of action, i.e. an action which reflects a moral value (e.g honesty), and an unethical behaviour, which reflects the possibility to deviate from the normative moral value for the sake of self-interested gain (e.g. self-interested lying). In Gunia et al. (2012) *contemplation* – defined as "individually conducted moral reasoning" – and *conversation* – defined as "social contemplation" – produce a positive influence on ethical decisions in an honesty vs lying game. This evidence is confirmed by other experiments on honesty, in which the decision to tell the truth appears to be favoured by longer deliberation times. This has been interpreted considering that honest behaviours presuppose self-control and hence the capacity to overcome the automatic self-serving response to situations in which cheating represents an advantageous opportunity to decision-maker. In this sense, "honesty requires time": subjects pushed to a quick answer tend to cheat more with respect to subjects who make decisions without time limits (Shalvi et al., 2012)[2]

[2]These findings may appear in conflict with the so-called Social Heuristics Hypothesis (Rand et al., 2012, 2014; Rand, 2016)), which on the ground of Dual Process Theory demonstrates that - in the context of social dilemma decisions (one-shot, anonymous cooperation games) - fast intuition is relatively more cooperative than deliberative responses (Evans and Stanovich, 2013; Alos-Ferrer and Strack, 2014). However, the types of decisions involved here are substantially different, and does not imply strong negative emotion (see Gunia et al., 2012 and Greene, 2014)..

For criticisms of findings and methodology of SHH see Tinghög et al. (2013), Tinghög et al. (2016). Dual Process Theory has been applied to risk-taking decisions, showing that the *reflection effect* of Prospect Theory is associated to automatic reactions, with "time pressure" increasing risk aversion in the case of gains and a reducing loss aversion (Kirchler et al., 2016).

Our hypothesis is that risk-taking for others can be assimilated to right-wrong decisions, with the moral right option made opaque by the frame of risk-decision-for-other in the initial situation. We hence expect that longer deliberation time and peers' discussion (representing ethic meetings in our case) favour a significant decrease in risk-taking over others. Our hypothesis rests on the consideration that, in risk-taking for others, reflection and peer-discussion enhance the "cognitive awareness of relevant moral values"(Gunia et al., 2012), so that the relevant moral norm ("do not risk over others") becomes salient to decision makers, compensating for the consolidated selfish attitude, which, as mentioned above, represents a rational criteria of decision in risk-taking on behalf of others. In this way we also contribute to the debate around the Social Heuristics Hypothesis (Rand et al., 2012, 2014; Rand, 2016), not with the purpose of rejecting it, but to re-integrate it within a wider conception of pro-sociality, in which deliberation and discussion play an active role in re-framing the decision situation, through a change in individual beliefs, so that the decision-maker, at the end of the process, acquire a more complete awareness of the consequences of the decision and of the value of the decision itself, and can adhere to previously unacknowledged moral norms.

3.3 Experimental design

The design of our experiment involves three phases: 1) the "social preference phase", in which we control for the preexistence of social preference; 2) the "decision phase", in which the risk-taking decision is made; 3) the "belief phase", in which we elicit beliefs (Charness and Dufwenberg, 2006) and collect responses to the short version of the big-five questionnaire (John et al., 1991, 2008). Phases 1 and 3 are common to all the treatments. Treatments' variations are hence applied only to phase 2 with the purpose ti identify the effect of ethics-meetings.

At their arrival, participants were randomly assigned to one of the two rooms set up for the experiment, with 14 computer-stations suited to guaranty anonymity. The first phase is a lottery which, following (Bartling et al., 2009), we run in order to control for

Table 3.1: Social Preference Choice

		X	**Y**
Line 1	*You*	2	2
	Your partner	2	1
Line 2	*You*	2	3
	Your partner	2	1
Line 3	*You*	2	2
	Your partner	2	4
Line 4	*You*	2	3
	Your partner	2	5

several types of social preferences. Each subject was exposed to 4 decisions in which she had to choose how to allocate payoffs between herself and another subject, randomly and anonymously paired to her in the same room. Everyone had to choose among allocation X and Y (see Table 3.1). At the end of the second phase, only one couple in the room (i.e. the winner and his partner) and one decision line was selected for payment.

The second phase consisted in a simple two-option choice, "*Left*" and "*Right*", which implied the roll of a six-faces dice. The result of this dice roll implied different payments for both the decision-maker (type B participant) and the passive receiver (type A participant). Payoffs are summarized in the Table 3.2. Each participant in each room knew to be randomly and anonymously paired to another subject in the other room. Participants also knew that participants in the other room were performing the same decision under the same conditions. Therefore, each participant in each room had to decide between "*Left*" and "*Right*" without knowing their role. Only at the very end of the experiment, all the participants in one room were assigned the role of B and, respectively, all the participants in the other room were assigned the role of A, according to the drawing of a card from a card deck (the first room picking a red card was assigning the B status). See the English translation of the instructions in the Appendix.

Participants in every treatments had to make the same choice implying the same payoffs table. To identify the effects of deliberation and ethics discussions we vary across treatment the time and the conditions of decision-making. In each session, the same

Table 3.2: Rolling the Die: Left or Right Choice

	LEFT		RIGHT	
Dice result	**A**	**B**	**A**	**B**
=1	6	16	0	6
≠1	0	6	6	6

If the decision maker chooses "Right" and 1 comes out, she gets 6 Euro and the recipient 0 Euro. If the decision maker chooses "Right", and a number different from one comes out, she gets 6 Euro and the recipient 6 Euro.
On the other hand, if she chooses "Left" and 1 comes out, she gets 16 Euro and the recipient 6 Euro. If she chooses "Left" and a number different from one comes out she gets 6 Euro and the recipient 0 Euro.

treatment was simultaneously performed in the two rooms. In BT, after the experimenter had firstly read the instrunctions aloud, participants were given only four minutes in order to let them fully understand the instruction and ask clarification questions to the experimenter. In the RT, after the first 4 minutes for understanding the instruction, participants were given 4 additional minutes to think alone and in complete silence.

In the EMT, participants were asked to discuss in groups, for 4 minutes (after the first 4 minutes for understanding the instruction) about the consequences of their decisions both for As and Bs and about their personal intentions concerning the decision at stake. In particular, in *Ethics 2* treatment participants were matched in pairs for a face-to-face interaction. In the *Ethics Chat* treatment, participants were matched in pairs but had to discuss through a chat, allowing us to track the content of conversations and to check whether there are differences in decisions due to communication mode. In the *Ethics 3* treatment, participants were instead matched in a triad for a face-to-face interaction in order to check for differences due to group size. In all EMT treatments, after the 4 minutes of discussion, participants were asked to turn back to their screen and make their choice individually. See Table 3.3 for an overview of treatments.

It is important to notice that ethics meetings could be considered linked to a risk of *experimenter's demand effect*, which is common to experimental studies aimed at investigating the effect of normativity and moral framing on decision making. In this regard, it must be considered that as far as a norm is explicitly concerned it is exactly the "demand effect" of that norm which is at stake. In our experiment, by organizing "ethic meetings" among decision-makers, we make the group of discussion internal to the

Table 3.3: Second Phase: Treatment Overview

Treatment	Time		Decision
Baseline (BT)	4 minutes*		Left or Right choice
Reflection (RT)	4 minutes*	4 minutes alone	Left or Right choice
Ethics- Meetings (EMT)			
Ethics 2	4 minutes*	4 minutes discussion in couple	Left or Right choice
Ethics 3	4 minutes*	4 minutes discussion in a group of three	Left or Right choice
Ethics Chat	4 minutes*	4 minutes discussione in couple through a chat	Left or Right choice

**Clarifications questions to the experimenter are allowed*

game. Consequently, the normative definition of what is appropriate to do emerges from interaction among subjects, who, at the same time, have an individual (selfish) interest in deviating from the norm itself. In this way, not only we can avoid the problem of a normative (demand) effect as produced from outside the decision-situation, but also we can analyse factors which determine (from within) the formation of the norm through belief change.

In the third phase, to assess whether the effect of ethic meetings over risk-taking decisions was mediated by expectations about others' decisions (Bicchieri, 2005) or by guilt aversion (Charness and Dufwenberg, 2006), we asked participants to specify their beliefs concerning: 1) the percentage of subjects choosing left in their same room ("peers belief"); 2) the percentage of subjects choosing left in the other room ("first order belief"); 3) the average response to question 2 in the other room ("second order belief"). Payments to this phase were determined by a lottery selecting only one of the three questions and by rewarding who had answered correctly to the selected question, with a 10% of tolerance. A questionnaire with big-five questions, and relevant personal information

(sex, age, years attending university) concluded the experiment[3].

The experiment took place at the "Laboratorio di Economia Sperimentale" of University of Pisa on January and May 2017. We conducted 16 sessions, each involving either 28, 24 or 20 participants, for a total of 412 participants invited from a pool of more than 1500 registered students from every departments of Pisa University. No one could take part to more than one session. Including the show-up fee of 5 €, the average pay was 10,90 €. In total, we run 2 sessions performing the *Baseline* treatment (BT), 2 sessions performing the *Reflection* treatment (RT), and 12 performing different types of the *Ethics-Meetings* treatments (EMT). In particular, we conduct 3 sessions with ethics-meetings composed of two persons (*Ethics Group 2*), 4 sessions with ethics-meetings composed of two persons communicating through a chat (*Ethics Chat)*, and 5 sessions with ethics-meetings composed of three persons (*Ethics Group 3).*

3.3.1 Hypotheses

The main hypotesis we tested was that participants in the EMT would have chosen *"Left"* less often than participants in the treatments without ethics discussion. Moreover, we made the hypothsis that also RT would have reduced the percentage of participants playng *"Left",* with respect to the BT. Therefore we expected that:

$$\%Left\ Choice_{BT} > \%Left\ Choice_{RT} > \%Left\ Choice_{EMT}$$

We also suppose that ethics meetings have an effect on risk-taking decisions for others by changing the beliefs concerning a) what other peers (in the same room) will choose in the decision-making situation; b) what others, outside the room of peers (namely, participants in the other room) would choose if asked to play as the decision makers; c)

[3]More precisely we asked the following questions:

1. Out of 10 participants, how many participants do you believe have played left in this room?
2. Out of 10 participants, how many participants do you believe have played left in the other room?
3. The same questions 2 has been asked to the participants in the other room. What do you think is the average answer to that question?

what the passive receivers of the consequences of the decision expect from the decision-maker.

3.4 Results

3.4.1 Univariate

In line with our hypothesis that ethics meetings favour more ethical (pro-socially responsible) choices, we observe that participants in the EMT showed the lowest percentage of "left" decisions (26%). Moreover, subjects allowed to reflect by themselves in the RT opted for left decisions less often (35%) than subjects in the BT (50%) (see Table 3.4), so confirming also the expectation that deliberation have a positive effect on ethical choices. These effects (which account for correlation at group level) are marginally significant when considering the difference between BT and RT (i.e. -0.14 with a p-value=0.063), while is significant when considering the difference between RT and EMT (i.e. -0.089 with p-value=0.051). In particular, within the EMT treatments, the largest effect is observed when participants discussed through a chat (i.e. -0.149 with p-value=0.032).

Interesting to note, in line with Schram and Charness (2015), who find that females are more likely to follow advices, we do find a stronger effect of ethics meetings when the group only comprises females. More precisely, when the group comprises females only, the average effect is significantly higher (i.e. -0.21 with a p-value=0.031). This effect is particularly large when the discussion occurs through a chat (i.e. -0.28 0.027). Instead, we do find a large effect of reflection in groups made of all males (i.e. -0.24 with a p-value=0.042). For mixed groups, the average effects is something in between these effects for females and males group (results are not reported).

But what are the drivers of these differences in choices across treatments? In which way ethics-meetings mold the way in which people think about their decision? As anticipated, in order to test these effects, we elicited beliefs of our participants. Indeed, if we compare the data concerning beliefs in each treatment, we can see that there are striking differences across treatments (Table 3.5). In particular, we can notice that in

Table 3.4: Choice side across treatments

Treatment	Indep Obs	Share Left		Difference	*p-value*
Baseline	56	0.5			
Reflection	56	0.35	**Reflection vs Baseline**	-0.14	0.063
Ethics (Total)	130	0.26	**Ethics vs Reflection**	-0.089	0.051
Ethics 2	42	0.27	*Ethics Group 2 vs Reflection*	-0.083	0.173
Ethics 3	40	0.31	*Ethics Group 3 vs Reflection*	-0.049	0.284
Ethics Chat	48	0.21	*Ethics Chat vs Reflection*	-0.149	0.032

Female Groups (i.e. all female)					
Treatment	Indep Obs	Share Left		Difference	*p-value*
Baseline	34	0.5			
Reflection	29	0.45	**Reflection vs Baseline**	0.05	0.344
Ethics (Total)	30	0.24	**Ethics vs Reflection**	-0.21	0.031
Ethics 2	10	0.30	*Ethics 2 vs Reflection*	-0.14	0.199
Ethics 3	5	0.33	*Ethics 3 vs Reflection*	-0.11	0.315
Ethics Chat	15	0.17	*Ethics Chat vs Reflection*	-0.28	0.027

Male Groups (i.e.all males)					
Treatment	Indep Obs	Share Left		Difference	*p-value*
Baseline	22	0.5			
Reflection	27	0.259	**Reflection vs Baseline**	-0.24	0.042
Ethics (Total)	22	0.265	**Ethics vs Reflection**	0.005	0.520
Ethics 2	7	0.367	*Ethics 2 vs Reflection*	0.097	0.307
Ethics 3	7	0.19	*Ethics 3 vs Reflection*	-0.06	0.315
Ethics Chat	8	0.25	*Ethics Chat vs Reflection*	-0.009	0.478

EMT treatments individuals always expect that participants - both in the same and in the other room - will choose "left" less frequently. For example, we observe a clear reduction in the share of people believed to play left in the same room ("Share Same Room"), falling from 59,46% of the BT, to 52,68% of the RT, to 41% of the EMT treatments. These differences are always economically and statistically significant. For example the difference in "Share Same Room" is about 12% lower in EMT than in RT with a p-value=0.000. This result hold for all types of EMT treatments, which have an analogous effect on all types of beliefs. For example, if we look at the first order beliefs ("Share other room"), we observe that EMT is about 12% lower than RT with a p-value=0.000, while if we look at expectations about others expectations ("Other Room Beliefs") EMT is 13% lower than RT with a p-value=0.000. It must be underlined that the larger shift in beliefs happens in the passage from RT to EMT, suggesting that only EMT is effective in changing decision making attitude towards the choice at stake.

These results suggests that ethics meetings have a strong effect in shifting participants expectation concerning the behaviour of their own peer group, from one more self-interested (or group-interested) vision to one more pro-social and ethical. It seems that, under the light of ethic meetings, subjects have more positive expectation concerning the behaviours that others – and in particular the receiver of the consequences of the decisions – would have adopted in the same decision situation. More precisely, participants in the EMT evaluates others opinion as less negative than how they perceive it in the BT and in the RT, and this is compatible with the hypothesis that the effect of ethic meetings over risk-taking for others – and in particular the observed reduction in the "Left" choices in EMT – depends on the circumstance that the decision-makers try to correspond to what they believe the others expect from them.

3.4.2 Multivariate

In order to investigate the possible determinants of the left choices and to size the effect of belief change on risk-taking for other across treatments we run a logit regression. The dependent variable is the individual choice (i.e. a dummy variable equal to 1 if individual

Table 3.5: Beliefs about Choice Side across treatments

Treatment	Indep Obs	**Share Same room**		Difference	*p-value*
Baseline	56	0.59			
Reflection	56	0.53	**Reflection vs Baseline**	-0.06	0.078
Ethics (Total)	130	0.41	**Ethics vs Reflection**	-0.12	0.000
Ethics Group 2	42	0.44	*Ethics 2 vs Reflection*	-0.09	0.029
Ethics Group 3	40	0.41	*Ethics 3 vs Reflection*	-0.12	0.006
Ethics Chat	48	0.40	*Ethics Chat vs Reflection*	-0.13	0.001

Treatment	Indep Obs	**Share other room**		Difference	*p-value*
Baseline	56	0.56			
Reflection	56	0.55	**Reflection vs Baseline**	-0.01	0.372
Ethics (Total)	130	0.43	**Ethics vs Reflection**	-0.12	0.000
Ethics Group 2	42	0.45	*Ethics 2 vs Reflection*	-0.10	0.018
Ethics Group 3	40	0.43	*Ethics 3 vs Reflection*	-0.12	0.004
Ethics Chat	48	0.42	*Ethics Chat vs Reflection*	-0.13	0.001

Treatment	Indep Obs	**other room belief**		Difference	*p-value*
Baseline	56	0.59			
Reflection	56	0.57	**Reflection vs Baseline**	-0.02	0.289
Ethics (Total)	130	0.46	**Ethics vs Reflection**	-0.13	0.000
Ethics 2	42	0.48	*Ethics 2 vs Reflection*	-0.11	0.002
Ethics 3	40	0.44	*Ethics 3 vs Reflection*	-0.13	0.000
Ethics Chat	48	0.45	*Ethics Chat vs Reflection*	-0.12	0.000

played left) while the independent variables are individuals' big five characterizations, their social preferences, as well as their belief concerning the percentage of participants choosing left. In particular, to avoid multicollinearity problems among first order and second order belief variables, we create a new variable "Beliefs Difference" which measures the difference between the "Share Other Room" and "Other Room Belief" variables (see Table3.6).

However, as the traditional tests of the equality of coefficients across groups in non-linear regressions are complicated by identification problems (i.e. the residual variation confound the magnitude of the effects, see Long, 2009), we report standardized coefficients. In particular, we notice that the main drivers across treatments of left choices are the beliefs variables. In particular, the effect of expectations about choices of people in the same room is always positive and statistically significant. For example, one standard deviation increase in the "Share same room" will increase the probability of choosing left by 1.198 in the BT. In the ethics treatments this effect although smaller is still at play. This suggests that a sort of *peer-effect* is generally at stake. Therefore, if the decision-maker believes that more peers are going to opt for "Left", it is more likely that she also conform to the same decision, trying to get as much as possible from the decision situation. These results are in line with the effects of empirical expectation as studied in Bicchieri (2005); Bicchieri and Xiao (2009).

On the other hand, the "Beliefs difference" coefficients suggest that a more complex *justification effect* involves the relationship between how the decision maker views others outside the room of peers and what she thinks about how others outside the room of peers view her and her peers, in relation to the decision situation. This effect is positive and significant only for the BT and for one treatment among the EMT treatments (i.e. *Ethics 2*). Since "Beliefs difference" grows when "Share Other Room" is larger than "Other Room Belief", it is relatively more likely that the decision maker chooses left if she thinks that the expectations of other room participants are lower and/or she believes that more participants in the other room are going to play left. In general terms, if I believe that the others will misbehave more than what I believe they expect from me,

I may feel justified to misbehave. On the contrary, if I believe that the others will behave more ethically, than what I believe they expect from me, I may be more willing to opt for the ethical choice. In this sense, we can consider that the above mentioned justification effect goes together with a sort of guilt aversion in the case of risking for other. Moreover, if we notice that this effect slightly reduces from BT to EMT, we can conclude that ethic meeting reduces risking over others also by reducing justification effect and/or increasing guilt aversion.

3.5 Conclusions

The results confirm the initial hypothesis that decisions concerning risk-taking for others can be interpreted as analogous to right-wrong decisions. Deliberation and peers discussion play a crucial role in this context, making the decision-maker aware of relevant consequences and of moral norms and values, not considered in the initial decision condition. In this respect, ethic meetings represent a potential effective instrument to reduce self-justification among financial agents, making the expectations (and benefits) of others who are mere receivers of the consequences of risk-taking decisions salient to the decision-maker. The observed reduction in risk-taking over others suggests the opportunity to further investigate the role of ethic-meetings in actual decision contexts, as a method to sustain a shift from the prevailing risk-seeking behaviours to a more responsible bank culture.

Table 3.6: Logit: Choice Left

	Baseline	*Reflection*	*Ethics 2*	*Ethics 3*	*Ethics Chat*
Extraversion	0.3511	0.4661	-0.1308	0.1144	-0.2974
	(0.480)	(0.594)	(0.298)	(0.266)	(0.326)
Agreeableness	-0.6353	-0.4646	-0.4336*	-0.1830	0.3338
	(0.548)	(0.477)	(0.263)	(0.222)	(0.454)
Conscientiousness	0.2221	-0.2912	-0.6230*	0.4613	-0.1250
	(0.524)	(0.471)	(0.341)	(0.327)	(0.346)
Neuroticism	0.2682	-0.1817	-0.0409	0.1675	0.1685
	(0.450)	(0.453)	(0.459)	(0.292)	(0.342)
Openness	0.2177	0.0016	0.0893	0.1637	0.1789
	(0.522)	(0.513)	(0.280)	(0.202)	(0.299)
Share same room	1.1982**	1.4515***	0.9744***	0.7079***	0.8013***
	(0.501)	(0.459)	(0.293)	(0.238)	(0.302)
Beliefs Difference	0.8145*	0.2388	0.7183***	0.2126	0.2573
	(0.436)	(0.421)	(0.257)	(0.507)	(0.325)
Prosocial	-0.6789*	-0.1224	-0.5494	-0.1743	-0.1934
	(0.394)	(0.321)	(0.361)	(0.258)	(0.371)
Costly prosocial	-0.0408	0.1878	-0.4818*	-0.6492**	-0.2392
	(0.406)	(0.428)	(0.266)	(0.288)	(0.313)
Envy	-0.6424*	0.0404	-0.1581	-0.1526	-0.6967**
	(0.384)	(0.444)	(0.325)	(0.264)	(0.296)
Costly Envy	-0.1975	0.3155	-0.5777	-0.1212	0.0422
	(0.377)	(0.366)	(0.359)	(0.310)	(0.305)
Female	-0.5156	0.5584	-0.0812	-0.0268	-0.0910
	(0.559)	(0.405)	(0.326)	(0.280)	(0.348)
Constant	-0.0632	-1.0095***	-1.3692***	-1.0197***	-1.7012***
	(0.358)	(0.382)	(0.367)	(0.215)	(0.360)
ll	-25	-27	-36	-59	-41
N	56	56	84	120	96

*p¡0.10,** p¡0.05, ***p¡0.01

Appendix: Second phase instructions

In this game you have to choose between two options: right or left. Both these options will imply the rolling of a six-faces dice. The dice result will have consequences on the money that will be paid to you and to another participant in the other room, at the end of the experiment.

An ID number will be randomly assigned to each participant. According to this number each participant in one room will be automatically matched to another participant in the other room. The number guarantees anonymity during the experiment and afterwards. Participants in one room will be assigned the role of type A player; while participants in the other room will be assigned the role of type B player.

At the end of the experiment, in each room, we will pick a card from a card deck. Participants in the room in which a red card will be picked (if two red cards are picked we go on with another draw) will be assigned the role of B player; the role of A player will be assigned to all participants in the other room. Payments will be determined in both rooms according to B choices, in the selected room.

Each participant in this room have to decide, only once, whether to play the right option or the left option. Participants in the other room will do the same.

In case you choose right and you are assigned the B role:

- If the dice result is 1: you will get 6 and the A player will get 0;
- If the dice result is different from 1: you will get 6 and the A player will get 6.

In case you choose left and you are assigned the B role:

- If the dice result is 1: you will get 16 and the A player will get 6;
- If the dice result is different from 1: you will get 6 and the A player will get 0.

In case you are assigned the A role the payment you will get will depend on the choice made by the player in the other room.

The payoffs are summarized in the following table:

	LEFT		RIGHT	
Dice result	**PAYOFF FOR A**	**PAYOFF FOR B**	**PAYOFF FOR A**	**PAYOFF FOR B**
NUMBER 1	6	16	0	6
NUMBER DIFFERENT FROM 1	0	6	6	6

Before choosing, we invite you to discuss for 4 minutes with the participant seated next to you "the consequences that your decision will have on the A player and on the B player" (but only with him/her). You can also discuss with him/her your intentions regarding the decision whether to play right or left. At the end of the discussion you will have to make your choice, right or left. Notice that this final choice will be private and anyone, including your discussion mate, can see it.

At the end of the experiment, the dice result and the result of the first phase of the experiment will be displayed on your screen. You will eventually be asked to fill a questionnaire.

We remind you that your participation will remain anonymous to other participants and to the experimenter. Your payment will be implemented by referring to the randomly assigned ID number which will appear on your screen.

You have now 4 minutes to read again carefully these instructions. During this time, if you have questions please rise your hand and we will personally answer you. After these 4 minutes, you can discuss with the participant seated next to you for other 4 minutes.

Thanks for your participation.

Bibliography

Achtziger, A., Alós-Ferrer, C., and Wagner, A. K. (2016). The impact of self-control depletion on social preferences in the ultimatum game. *Journal of Economic Psychology*, 53:1–16.

Alos-Ferrer, C. and Strack, F. (2014). From dual processes to multiple selves: Implications for economic behavior. *Journal of Economic Psychology*, 41:1–11.

Andersson, O., Holm, H. J., Tyran, J.-R., and Wengström, E. (2013). Risking other people s money.

Andersson, O., Holm, H. J., Tyran, J.-R., and Wengström, E. (2014). Deciding for others reduces loss aversion. *Management Science*, 62(1):29–36.

Aoki, M. (2007). Endogenizing institutions and institutional changes. *Journal of Institutional Economics*, 3(01):1–31.

Aoki, M. (2011). Institutions as cognitive media between strategic interactions and individual beliefs. *Journal of Economic Behavior & Organization*, 79(1):20–34.

Awrey, D., Blair, W., and Kershaw, D. (2013). Between law and markets: Is there a role for culture and ethics in financial regulation? *Delaware Journal of Corporate Law (DJCL)*, 38:191.

Bacharach, M. (2006). *Beyond individual choice: teams and frames in game theory.* Princeton University Press.

Balliet, D. (2010). Communication and cooperation in social dilemmas: A meta-analytic review. *Journal of Conflict Resolution*, 54(1):39–57.

Bartling, B., Fehr, E., Maréchal, M. A., and Schunk, D. (2009). Egalitarianism and competitiveness. *The American Economic Review*, 99(2):93–98.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms.* Cambridge University Press.

Bicchieri, C. and Xiao, E. (2009). Do the right thing: but only if others do so. *Journal of Behavioral Decision Making*, 22(2):191–208.

Binmore, K. (2015). Institutions, rules and equilibria: a commentary. *Journal of Institutional Economics*, 11(03):493–496.

Binmore, K. G. (1994). *Game theory and the social contract: playing fair*, volume 1. MIT press.

Binmore, K. G. (1998). *Game theory and the social contract: just playing*, volume 2. MIT press.

Blount, S. (1995). When social outcomes aren't fair: The effect of causal attributions on preferences. *Organizational behavior and human decision processes*, 63(2):131–144.

Bohnet, I., Greig, F., Herrmann, B., and Zeckhauser, R. (2008). Betrayal aversion: Evidence from brazil, china, oman, switzerland, turkey, and the united states. *The American Economic Review*, 98(1):294–310.

Bolton, G., Ockenfels, A., et al. (2008). Risk taking and social comparison. a comment on "betrayal aversion: Evidence from brazil, china, oman, switzerland, turkey, and the united states". *American Economic Review*, 100(1):628–633.

Bolton, G. E. and Ockenfels, A. (2000). Erc: A theory of equity, reciprocity, and competition. *American economic review*, pages 166–193.

Bowles, S. (1998). Endogenous preferences: The cultural consequences of markets and other economic institutions. *Journal of economic literature*, pages 75–111.

Bowles, S. (2004). *Microeconomics: behavior, institutions, and evolution.* Princeton: Princeton University Press.

Bowles, S. (2008). Policies designed for self-interested citizens may undermine" the moral sentiments": Evidence from economic experiments. *science*, 320(5883):1605–1609.

Bowles, S. and Polania-Reyes, S. (2012). Economic incentives and social preferences: substitutes or complements? *Journal of Economic Literature*, pages 368–425.

Burks, S. V. and Krupka, E. L. (2012). A multimethod approach to identifying norms and normative expectations within a corporate hierarchy: Evidence from the financial services industry. *Management Science*, 58(1):203–217.

Chang, R. (1997). *Incommensurability, incomparability, and practical reason.* Harvard University Press.

Charness, G. and Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, 74(6):1579–1601.

Charness, G. and Rabin, M. (2002). Understanding social preferences with simple tests. *Quarterly journal of Economics*, pages 817–869.

Cohn, A., Fehr, E., and Maréchal, M. A. (2014). Business culture and dishonesty in the banking industry. *Nature*, 516(7529):86–89.

Cohn, A., Fehr, E., and Maréchal, M. A. (2015). A culture of gambling? evidence from the banking industry.

Cole, S., Kanz, M., and Klapper, L. (2015). Incentivizing calculated risk-taking: Evidence from an experiment with commercial bank loan officers. *The Journal of Finance*, 70(2):537–575.

Crespo, R. F. (2007). 'practical comparability'and ends in economics. *Journal of Economic Methodology*, 14(3):371–393.

Crespo, R. F. (2016). Aristotle on agency, habits and institutions. *Journal of Institutional Economics*, pages 1–18.

Davis, J. B. (2003). *The theory of the individual in economics: Identity and value.* Routledge.

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of personality and Social Psychology*, 18(1):105.

Ditto, P. H., Pizarro, D. A., and Tannenbaum, D. (2009). Motivated moral reasoning. *Psychology of learning and motivation*, 50:307–338.

Evans, J. S. B. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.*, 59:255–278.

Evans, J. S. B. and Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, 8(3):223–241.

Falk, A., Fehr, E., and Fischbacher, U. (2003). On the nature of fair behavior. *Economic Inquiry*, 41(1):20–26.

Fehr, E. and Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and human behavior*, 25(2):63–87.

Fehr, E. and Gächter, S. (2000). Fairness and retaliation: The economics of reciprocity. *The journal of economic perspectives*, 14(3):159–181.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly journal of Economics*, pages 817–868.

Frey, B. S. (1997). *Not just for the money: An economic theory of personal motivation.* Edward Elgar Pub.

Frey, B. S. and Jegen, R. (2001). Motivation crowding theory. *Journal of economic surveys*, 15(5):589–611.

Gallotti, M. and Frith, C. D. (2013). Social cognition in the we-mode. *Trends in cognitive sciences*, 17(4):160–165.

Gilbert, M. (1989). *On social facts*. Princeton University Press.

Gilbert, M. (2007). Searle and collective intentions. In *Intentional acts and institutional facts*, pages 31–48. Springer.

Gilbert, M. (2013). *Joint commitment: How we make the social world*. Oxford University Press.

Gintis, H. (2007). The evolution of private property. *Journal of Economic Behavior & Organization*, 64(1):1–16.

Gintis, H. (2009). *The bounds of reason: game theory and the unification of the behavioral sciences*. Princeton University Press.

Gintis, H., Bowles, S., Boyd, R., and Fehr, E. e. (2005). *Moral sentiments and material interests: The foundations of cooperation in economic life*, volume 6. MIT press.

Gneezy, U. (2005). Deception: The role of consequences. *The American Economic Review*, 95(1):384–394.

Gold, N. (2012). Team reasoning, framing and cooperation. *Evolution and rationality: decision, cooperation and strategic behaviour*, pages 185–212.

Gold, N. and Sugden, R. (2007). Collective intentions and team agency. *The Journal of Philosophy*, pages 109–137.

Greene, J. (2013). *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*. Penguin.

Greene, J. D. (2007). The secret joke of kant's soul. *Moral Psychology: Historical and Contemporary Readings*, pages 359–372.

Greene, J. D. (2009). Dual-process morality and the personal/impersonal distinction: A reply to mcguire, langdon, coltheart, and mackenzie. *Journal of Experimental Social Psychology*, 45(3):581–584.

Greene, J. D. (2014). Beyond point-and-shoot morality: Why cognitive (neuro) science matters for ethics. *Ethics*, 124(4):695–726.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., and Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, 44(2):389–400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., and Cohen, J. D. (2001). An fmri investigation of emotional engagement in moral judgment. *Science*, 293(5537):2105–2108.

Greif, A. (2006). *Institutions and the path to the modern economy: Lessons from medieval trade*. Cambridge University Press.

Greif, A. and Kingston, C. (2011). Institutions: Rules or equilibria? In *Political economy of institutions, democracy and voting*, pages 13–43. Springer.

Guala, F. and Hindriks, F. (2014). Institutions, rules, and equilibria: a unified theory. *Journal of Institutional Economics*, pages 1–22.

Guala, F. and Hindriks, F. (2015). A unified social ontology. *The Philosophical Quarterly*, 65(259):177–201.

Guiso, L., Sapienza, P., and Zingales, L. (2015). The value of corporate culture. *Journal of Financial Economics*, 117(1):60–76.

Gunia, B. C., Wang, L., Huang, L., Wang, J., and Murnighan, J. K. (2012). Contemplation and conversation: Subtle influences on moral decision making. *Academy of Management Journal*, 55(1):13–33.

Habermas, J. (1990). *Moral consciousness and communicative action*. MIT press.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, 108(4):814.

Haidt, J. (2003). The moral emotions. *Handbook of affective sciences*, 11:852–870.

Haidt, J. (2007). The new synthesis in moral psychology. *science*, 316(5827):998–1002.

Ham, J. and van den Bos, K. (2010). On unconscious morality: The effects of unconscious thinking on moral decision making. *Social Cognition*, 28(1):74–83.

Harman, G., Mason, K., and Sinnott-Armstrong, W. (2011). Moral reasoning. *The handbook of moral psychology. Oxford University Press, Oxford.*

Hausman, D., McPherson, M., and Satz, D. (2016). *Economic analysis, moral philosophy, and public policy.* Cambridge University Press.

Hausman, D. M. and McPherson, M. S. (1993). Taking ethics seriously: economics and contemporary moral philosophy. *Journal of economic literature*, 31(2):671–731.

Hindriks, F. (2009). Constitutive rules, language, and ontology. *Erkenntnis*, 71(2):253–275.

Hodgson, G. M. (2006). What are institutions? *Journal of Economic Issues*, 40(1):1–23.

Hodgson, G. M. (2012). *From pleasure machines to moral communities: an evolutionary economics without homo economicus.* University of Chicago Press.

Hodgson, G. M. (2015). On defining institutions: rules versus equilibria. *Journal of Institutional Economics*, pages 1–9.

Jacob, P. (2014). Intentionality. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy.* Winter 2014 edition.

John, O. P., Donahue, E. M., and Kentle, R. L. (1991). The big five inventory: versions 4a and 54.

John, O. P., Naumann, L. P., and Soto, C. J. (2008). Paradigm shift to the integrative big five trait taxonomy. *Handbook of personality: Theory and research*, 3:114–158.

Kahane, G., Wiech, K., Shackel, N., Farias, M., Savulescu, J., and Tracey, I. (2012). The neural basis of intuitive and counterintuitive moral judgment. *Social cognitive and affective neuroscience*, 7(4):393–402.

Kahneman, D. (2011). *Thinking, fast and slow.* Macmillan.

Karpus, J. and Gold, N. (2016). Theory and evidence. *The Routledge Handbook of Philosophy of the Social Mind*, page 400.

Kirchler, M., Andersson, D., Bonn, C., JOHANNESSON, M., SORENSEN, E., Stefan, M., TINGHOG, G., and VASTFJALL, D. (2016). The effect of fast and slow decisions on financial risk taking.

Klooster, J. and Meyer, M. (2016). Ethical banking: the key concepts. Mimeo, Cambridge Judge Business School.

Kohlberg, L. (1969). Stage and sequence: The cognitive-developmental approach to socialization. In Gosling, editor, *Handbook of Socialization Theory and Research.* Rand McNally.

Krajbich, I., Bartling, B., Hare, T., and Fehr, E. (2015). Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nature communications*, 6.

Krupka, E. L. and Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, 11(3):495–524.

Levine, D. K. (1998). Modeling altruism and spitefulness in experiments. *Review of economic dynamics*, 1(3):593–622.

Lewis, D. (1969). *Conventions.* Harvard University Press.

Liu, B. S. and Ditto, P. H. (2013). What dilemma? moral evaluation shapes factual belief. *Social Psychological and Personality Science*, 4(3):316–323.

Lo, A. W. (2015). The gordon gekko effect: The role of culture in the financial industry. Technical report, National Bureau of Economic Research.

Long, J. S. (2009). Group comparisons in logit and probit using predicted probabilities. *Department of Sociology, University of Indiana. Unpublished manuscript.*

Moore, C. and Tenbrunsel, A. E. (2014). "just think about it"? cognitive complexity and moral choice. *Organizational Behavior and Human Decision Processes*, 123(2):138–149.

Morris, N. and Vines, D. (2014). *Capital failure: Rebuilding trust in financial services.* OUP Oxford.

Nichols, S. and Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition*, 100(3):530–542.

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *science*, 314(5805):1560–1563.

Paxton, J. M. and Greene, J. D. (2010). Moral reasoning: Hints and allegations. *Topics in cognitive science*, 2(3):511–527.

Paxton, J. M., Ungar, L., and Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science*, 36(1):163–177.

Piaget, J. (1965). *The moral judgement of the child.* Free Press.

Pollmann, M. M., Potters, J., and Trautmann, S. T. (2014). Risk taking by agents: The role of ex-ante and ex-post accountability. *Economics Letters*, 123(3):387–390.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American economic review*, pages 1281–1302.

Rand, D. G. (2016). Cooperation, fast and slow meta-analytic evidence for a theory of social heuristics and self-interested deliberation. *Psychological Science*, page 0956797616654455.

Rand, D. G., Greene, J. D., and Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, 489(7416):427.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., and Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature communications*, 5.

Ryan, R. M. and Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology*, 25(1):54–67.

Sánchez-Cuenca, I. (2007). A behavioural critique of searle's theory of institutions. In *Intentional Acts and Institutional Facts*, pages 175–189. Springer.

Sandel, M. J. (1982). *Liberalism and the Limits of Justice*. Cambridge University Press.

Schotter, A. (1981). *The economic theory of social institutions*. Cambridge University Press.

Schram, A. and Charness, G. (2015). Inducing social norms in laboratory allocation choices. *Management Science*, 61(7):1531–1546.

Searle, J. R. (1995). *The construction of social reality*. Simon and Schuster.

Searle, J. R. (2005). What is an institution. *Journal of institutional economics*, 1(1):1–22.

Searle, J. R. (2010). *Making the social world: The structure of human civilization*. Oxford University Press.

Searle, J. R. (2015). Status functions and institutional facts: reply to hindriks and guala. *Journal of Institutional Economics*, pages 1–8.

Shalvi, S., Eldar, O., and Bereby-Meyer, Y. (2012). Honesty requires time (and lack of justifications). *Psychological science*, 23(10):1264–1270.

Sugden, R. (1986). *The economics of rights, co-operation and welfare.* Basil Blackwell, 2004 2nd edition.

Sugden, R. (2015). On 'common-sense ontology': a comment on the paper by frank hindriks and francesco guala. *Journal of Institutional Economics*, 11(03):489–492.

Suter, R. S. and Hertwig, R. (2011). Time and moral judgment. *Cognition*, 119(3):454–458.

Tinghög, G., Andersson, D., Bonn, C., Böttiger, H., Josephson, C., Lundgren, G., Västfjäll, D., Kirchler, M., and Johannesson, M. (2013). Intuition and cooperation reconsidered. *Nature*, 498(7452):E1–E2.

Tinghög, G., Andersson, D., Bonn, C., Johannesson, M., Kirchler, M., Koppel, L., and Västfjäll, D. (2016). Intuition and moral decision-making–the effect of time pressure and cognitive load on moral judgment and altruistic behavior. *PloS one*, 11(10):e0164012.

Tomasello, M. (2009). *Why we cooperate.* MIT press.

Tomasello, M. (2010). *Origins of human communication.* MIT press.

Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, 28(05):675–691.

Tuomela, R. (1991). We will do it: An analysis of group-intentions. *Philosophy and Phenomenological Research*, 51(2):249–277.

Tuomela, R. (2005). We-intentions revisited. *Philosophical Studies*, 125(3):327–369.

Tuomela, R. and Miller, K. (1988). We-intentions. *Philosophical Studies*, 53(3):367–389.

Umphress, V. J., Lambert, M. J., Smart, D. W., Barlow, S. H., and Clouse, G. (1997). Concurrent and construct validity of the outcome questionnaire. *Journal of Psychoeducational Assessment*, 15(1):40–55.

Velleman, J. D. (2009). *The possibility of practical reason.* Michigan Publishing, University of Michigan Library.

Viskovatoff, A. (2003). Searle, rationality, and social reality. *American Journal of Economics and Sociology*, 62(1):7–44.

Wehinger, G. et al. (2012). Banking in a challenging environment: business models, ethics and approaches towards risks. *OECD Journal: Financial Market Trends*, 2:79–88.

Young, H. P. (1998). *Individual strategy and social structure: An evolutionary theory of institutions.* Princeton University Press.

Young, H. P., Noe, T., et al. (2012). The limits to compensation in the financial sector. Technical report, Department of Economics Oxford.

Zaibert, L. and Smith, B. (2007). The varieties of normativity: An essay on social ontology. In Tsohatzidis, S. L., editor, *Intentional acts and institutional facts*, pages 157–173. Springer.

Zaibert, L. A. (2003). Collective intentions and collective intentionality. *American Journal of Economics and Sociology*, 62(1):209–232.

Zaki, J. and Mitchell, J. P. (2013). Intuitive prosociality. *Current Directions in Psychological Science*, 22(6):466–470.

Zhong, C.-B. and Liljenquist, K. (2006). Washing away your sins: Threatened morality and physical cleansing. *Science*, 313(5792):1451–1452.

Zingales, L. (2015). Does finance benefit society? Technical report, National Bureau of Economic Research.