Informatica e diritto, XXXIX annata, Vol. XXII, 2013, n. 1, pp. 75-98

Cognitivizing "Norms". Norm Internalization and Processing

CRISTIANO CASTELFRANCHI*

1. OUR PERSPECTIVE AND CLAIMS

What we present is not an agent-based simulation work[1]; it is more a theoretical contribution to normative cognition, the "psychology" of norms in

* The Author is research associate at the Institute of Cognitive Sciences and Technologies, National Research Council of Italy (ISTC-CNR), "GOAL" Group, Theoretical Psychology Project, Rome. This work is the estended text of a talk given at the European University Institute in Fiesole for a WS on "Norm compliance", July 2010, https://sites.google.com/site/normcompliance2010/program. I thank the participants for the nice discussion. I am also in debt with Rosaria Conte (many years working together or in parallel on these issues), Luca Tummolini, Giulia Andrighetto (for specific contributions on these issues) and the other members of the GOAL group for the general framework and precious feedback.

[1] I am sorry to disappoint my reader, but my contribution is not a discussion of the literature (philosophical, sociological, psychological, and AI) on norms and their working. It is more a restatement of the main issues of our work on norms in the last 15 years, work that has significantly contributed to social simulation studies on norms (see LABSS work http://www.istc.cnr.it/group/labss; R. CONTE, G. ANDRIGHETTO, M. CAMPENNÌ (eds.), *Minding Norms. Mechanisms and Dynamics of Social Order in Agent Society*, Oxford Series on Cognitive Models and Architectures, New York, Oxford University Press, forthcoming, and to the Agent and MAS research (see, for example, NorMAS WS and its community, G. BOELLA, P. NORIEGA, G. PIGOZZI, H. VERHAGEN (eds.), *Normative Multi-agent Systems*, Dagstuhl Seminar Proceedings 09121, 2009; G. ANDRIGHETTO, G. GOVERNATORI, P. NORIEGA, L. VAN DER TORRE (eds.), *Normative Multi-agent Systems*, Dagstuhl Seminar Proceedings 12111, 2012; and the Agreement Technology EU Project, http://www.agreement-technologies.eu/wg2, and G. ANDRIGHETTO, C. CASTEL-

a cognitive science perspective, strongly based on the computational modeling of social minds and interactions, on the Artificial Intelligence (AI) model of cognitive autonomous agents, and artificial society.

Our general perspective[2] is the following one:

- Social phenomena are due to the agents' behaviors, but the agents' behaviors are due to the mental mechanisms controlling and (re)producing them.
- It is impossible to understand the efficacy and working of the norms (Ns) without a modeling of how Ns succeed in changing our control system (mind) and regulating from inside our behavior.
- How does a norm (N) change and work through the minds of the agents? How is it "represented"?
- Which are the proximate mechanisms underlying the normative behavior?
- What does it mean to "conform" to a N from a mental – not just a behavioral – point of view? What does it mean to "obey"?
- What kind of mental attitude the N "prescribes" to, and builds into the agents?

1.1. Norms are artifacts, tools for the manipulation and regulation of autonomous cognitive agents' conduct; agents that have their own internal goals and decision processes. This happens thanks to the manipulation of our goals and preferences/choices.

Can we model how N succeed in giving us goals and intentions? They are built for that.

1.2. In many organizational, anthropological, sociological views not only there is a very strong (if not exclusive) emphasis on "sanctions" as necessary and "definitional" for having "norms" (Section 5.), but there is an explicit or implicit view of Ns (of organization, of institutions) as aimed at, having the function of: creating constraints/binds on the agents' behaviors in order to obtain a given coordinated collective behavior ("order").

The other face of Ns is ignored: the purpose and function of "inducing" goals in people, of influencing them to do something: to intend to do something; a goal that was not at all in their mind (Section 2.1.).

FRANCHI (eds.), *Norms*, in Ossowski S. (ed.), "Agreement Technologies, Law", Governance and Technology Series 8, 2013.

[2] R. CONTE, C. CASTELFRANCHI, *Cognitive and Social Action*, London, UCL Press, 1995.

Ns are not aimed just at "pruning" possible actions, or "permitting" them; at blocking some possible choice or changing the evaluation by altering the expected outcomes (rewards) of the alternatives. They are also aimed at adding, creating new goals and alternatives.

1.3. In the aforementioned view, it seems (it is implicitly assumed) that: goals (and then intentions) of an agent are all "desires", are all endogenous; and we have just to cut some possible course of action by making some desire practically impossible or non convenient.

It is ignored the fact that "duties" are not "desires"; they are goals from a different source, with a different origin: they come from outside (exogenous)[3], they are imported, "adopted"; they are "prescriptions" and "imperatives" from another Agent (the group, the authority).

Society (and "super-Ego") does not only "block" us, but gives us new goals, shapes our motivation, induce us to do, to pursue, something that might have never been in our mind.

We need a different mind "architecture" not simply based on BDI (beliefs, desires, intentions) models[4].

1.4. A N is not just aimed at regulating our conduct, at inducing us to do or not to do a given action; it is aimed at inducing us to do that action for specific motives, with a given mental attitude. The ideal-typical Adhesion to a N is for an intrinsic motivation, for a "sense of duty", recognition of the authority, because it is right/correct to respect Ns...; and only sub-ideally one should respect for avoiding external or internal sanctions (see Section 3.). Also normative education goes in this direction (see Section 6.3.).

1.5. Ns have to be "impersonal" and depersonalized (and perceived as such) on both sides: the issuer's and the addressee's side. It is not a conflict between me and you; it is not "my" request (for me, for my desires, etc. for my personal will that you have to adopt); and it is not a request to "you". The message is:

- "I do not talk, monitor, sanction, in my name";
- "I am not addressing to you 'ad personam', but as an instance of a class, a member, a citizen, ... like any other in the same conditions". Also for that "You have no reasons for rebelling".

[3] However, see later about the internalization of the "authority", and internal moral imperatives.

[4] On the BDI models and logics see for example http://www.loa.istc.cnr.it/Files/bdi.pdf.

This really is a crucial point in the perception of Ns as Ns; thus it is something that must be signaled in some way (uniform, role symbols, specific documents, . . .) or at least contextually presupposed and assumed in the script. The transition from personal power and violence to formal power, from the prince's thugs to policemen, has been a fundamental historical evolution; psychologically and culturally complex and something sham and hypocrite.

1.6. As we said, Ns are social device controlling behaviors through minds but in a specific way; through a partial understanding. They require (for their existence and effectiveness) their explicit mental representation, their (partial) understanding and recognition "as Norms"; specific cognitive representations and motivational processes (*Cognitive Mediators*)[5]; differently from other social phenomena like social functions, that can be played by social actors even without understanding – and even less intending – them[6].

1.7. Ns have to build in us an "ought", a "duty", "you have to"; with a rather constrictive feeling, a negative "frame", an avoidance orientation (even when it elicits "you have to do this action"). And this "ought" is a non-technical "ought", not instrumental to and planned for a given outcome/goal.

This entails a process of Adopting without sharing the "instrumental" nature of the N, and without understanding/adopting its "function" or end. My "plan" is different from the authority's "plan".

Citizens are not real "cooperators" but "subjects". They have to "alienate" their own powers and products (see Section 6.).

1.8. N require different and complementary "roles" with their specific "minds" or "mental attitudes": the subject of the N, the watcher, the issuer (see Section 10.).

And we have to explain how and why a subject of the N also becomes a watcher and an (implicit) issuer of it (see Section 4.).

1.9. We will also argue against

– the reduction of Ns to sanction/incentives and utility;

[5] R. CONTE, C. CASTELFRANCHI, *Cognitive and Social Action*, cit.; R. CONTE, C. CASTELFRANCHI, *The Mental Path of Norms*, in "Ratio Juris", Vol. 19, 2006, n. 4, pp. 501-517.

[6] C. CASTELFRANCHI, *The Theory of Social Functions. Challenges for Multi-agent-based Social Simulation and Multi-agent Learning*, in "Journal of Cognitive Systems Research", 2001, n. 2, 2001, pp. 5-38, http://www.cogsci.rpi.edu/ ~ rsun/si-mal/article1.pdf.

- the reduction of Ns to reinforcement learning and automatisms;
- the reduction of Ns to mere hardwired impossibility to act. Ns are "norms" only if they presuppose/allow the possibility of "violation".

2. Norm Internalization

Ns are based on a specific process of Goal-adoption or better adhesion; since they have the nature of an "imperative". That is, they are aimed at being "obeyed" for specific motives: not for external rewards, not for benevolence, etc. but for the recognition of authority, role, values, ...

2.1. Goal-adoption

Ns induce new goal through "adoption". Goal-adoption is how an autonomous agent is not an isle but becomes social, or better pro-social[7]; that is s/he does something for the others; puts her/his autonomous goal-pursuing (intentional action), her/his cognitive machinery for that, and her/his powers and resources, into the service of the others and of their interests. How is this possible? Not only economically or evolutionary, but cognitively, that is from the point of view of the working of an autonomous, self-regulated, goal-driven system. What kind of mental representations and operations are needed?

How is it possible that the goal (need, desire, objective, request, order, ...) of another entity succeeds in regulating my own autonomous behavior? How such a goal is "imported" in my regulatory, purposive system?

What is needed is the architecture of a social agent able to import goals from outside (and to influence other agents by giving them goals and relying on him/her) but remaining "autonomous". S/he is able to arrive to intentions not only from her own endogenous "desires", but from imported goals.

Goal-adoption means that:

X believes that Y has the Goal that p (G_y p) and comes to have (and possibly pursue) the Goal that p (G_x p) just because he believes this.

$$(\text{Goal-adopt x y p}) =^{def} (\text{R-Goal x p (BEL x(Goal y p))})$$

[7] Not to be used as synonim of "altruistic", "benevolence", etc. (see below).

"I do something 'for' you" (which does not mean "benevolence"!); I want to realize this since and until you wants/needs this; because it is your goal. Not a trivial notion, to be defined and formalized[8].

2.2. Reasons for Goal-adoption

There are different kinds of Goal-adoption, motivated by different reasons.

a) *Terminal or Altruistic*: Adoption can (rarely) be "altruistic", that is disinterested, non motivated by, non instrumental to higher personal (non-adoptive) calculated advantages (goals);

b) *Instrumental*: Adoption can be instrumental to personal/private returns, part of a selfish plan; like in commerce, where:

> It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner, but from their regard to their own interest. We address ourselves, not to their humanity but to their self-love, and never talk to them of our own necessities but of their advantages[9].

In Adam Smith's perfect characterization of exchange in merely selfish terms it is clear that there is not benevolence at all; and that X has the goals to understand and realize the selfish goal of Y (that *per se* is indifferent – or bad – to X) only in order to satisfy (through Y's reciprocal adoption) his own selfish and personal goal. So having the goal to realize your goal (as what you like and because you like it) is not necessarily altruistic at all.

c) *Cooperative*: it can be instrumental to a personal advantage, but shared with the other: in view of a common goal (strict "cooperation"): X and Y depend on each other for one and the same goal.

One might consider (c) a sub-case of (b) (instrumental adoption) but actually the situation is significantly different. While in (b) it is rational to try to

[8] This formalization might also cover "imitation" ("Since y wants p, me too!"). Goal-adoption is not "doing the same", "doing like the other". It is doing something "for" the other, for realizing her Goal. A better formalization should make clear that:

$$(\text{Goal-adopt x y p}) =^{def} (\text{R-Goal x (OBTAIN y p) (BEL x(Goal y p))})$$

X has the Goal that *Y realizes her Goal*, that she achieves it. Where the realization of "OBTAIN" implies (Knows y p) = p & (Bel y p).

[9] A. SMITH, *An Inquiry into the Nature and Causes of the Wealth of Nations*, 1776.

cheat and defeat the other; in strict cooperation, where we need each other for realizing one and the same goal, to defeat would be self-defeating.

2.3. Goal-adhesion

A stronger form of G-Adoption is Adhesion:

when I adhere to your (implicit or explicit) "request" (of any kind: prey, favor, order, law, etc.).

In other words, you (Y) have the goal that I adopt your goal p, that I do something (action a of X) realizing that goal, and I adopt your goal p or of doing a, (also) because I know that you expects and wants so.

In Adhesion one of the reasons for Adopting the goal of the other is that the other wants so:

- She also has the (meta)goal that we adopt her goal;
- We Adopt her goal by adopting the meta-goal.

In a sense, there is a double level of adoption (a meta-adoption): I know and adopt your goal that I adopt.

Moreover, in case of Adhesion there is a (presupposed) agreement between X and Y about X's adoption, X doing something as desired by Y. Other forms of adoption (like help) can be unilateral, spontaneous, and even against Y's desire.

3. Not Only Prescribed Behaviors But Expected Mental Attitudes

The aim of a N is not just our behavior; for example, the norm is not satisfied by an accidental conformity.

Ns have very strange claims on our mind: they prescribe also a "mental attitude". A N wants to be followed for an internal mechanism reflecting it; for a goal of following Ns. Moreover, the objective of the N on my mind is that I do not adopt its "request" for whatever reason (higher-goal) (pity, friendship, agreement on the content, personal advantage, fear, . . .).

I have to Adhere for specific reasons and higher-goals: for an intrinsic motivation (no external rewards), for a non-instrumental goal of respecting the authority and its norms. This is "obedience".

This is the real difference between an "order", a "favor", a "prayer", . . . ; not just social, relational, or pragmatic aspects. In all cases there is a request about an action (or inaction) of yours, but you have to do that for different reasons; I ask you a specific mental attitude towards my expectation and me.

The "order" of a general should not be "obeyed" because of courtesy, sympathy, friendship, pity, agreement about the solution, fear, money, … but just because it is an "order" of the right person, this is its "ideal" working, its aim[10]. Analogously, N wants:

- my behavior; due to
- my goal; due to
- my adhesion; but,
- motivated by specific higher-goal.

Ns are *not* based on an explicit or implicit "agreement", acceptance of us; the obligation applies and impinges on us in any case; it does not depend on our consent. For that we are "subjects" of the N. Ns do not "ask" us to do something (where we can reject the commitment not only the action); Ns "order/impose" us to do something, and we could "not do" as prescribed but could not "have not to do"; we can refuse to do the action but we cannot cancel the duty.

There is an "acceptance" (which is presupposed in the issuing of norm and in its obedience and real violation) but in another sense. The normative regulation relies on the fact that we "recognize" the N as a norm, and thus "acknowledge" its authority, and treat it accordingly with such assumption, even while disobeying.

Let us give another example of why prescribed mental attitudes matter more than the behavior: the "claim" for the "recognition" of a "right". Later we will say something more about a "duty" attitude (Section 7.).

3.1. Interpersonal Rights

I call "interpersonal rights" those rights that are not necessarily established by some law and imposed by some authority. Let me consider their "psychology". Consider the mental attitudes of two agents when one is claiming a right, and the other is acknowledging it[11].

Suppose that, on a bus, I want you to give me your seat because I think (and I want you to acknowledge) that this is my right (suppose I am a pregnant woman, an old man, etc.); and suppose that there is no official norm

[10] C. CASTELFRANCHI, *Prescribed Mental Attitudes in Goal-Adoption and Norm-Adoption*, in "Artificial Intelligence and Law", Special Issue on Norms in MAS, Vol. 7, 1999, pp. 37-50.

[11] C. CASTELFRANCHI, *Formalising the Informal? Dynamic Social Order, Bottom-up Social Control, and Spontaneous Normative Relations*, in "Journal of Applied Logic", Vol. 1, 2003, n. 1-2, pp. 47-92.

or rule about this. My goal is that you leave your seat, but this is not sufficient. For example, if, for independent reasons (the bus arrived at your bus stop), you leave, your seat is free but my right has not been acknowledged. What is necessary here is Goal-adoption: you have to know and adopt my goal. But even this is not sufficient: I pretend much more from your mind: *I want you to adopt my goal with a specific mental attitude and for specific reasons* (higher-goals, motivations). You might leave your seat just out of pity: this is not "acknowledging my right"; you might do it out of love, sympathy, courtship: but this is not "acknowledging my right" either. You could do it out of fear or interest, because I am very strong and I am threatening you, or because I offered you 2 dollars: again, this is not "acknowledging my right".

Summing up, I do not want you to adopt my goal only. I want you to do this because you believe (agree) that this is my right, that my request/expectation is correct and you perform the action *in order to* respect rights[12].

Why should we be so interested in the mind behind the action, when seemingly what we practically need is that action? The truth is that we do not need only or mainly the required action. In the case of rights it is quite clear which are the very different social consequences of the different attitudes you have in adopting my goal. If you do it for pity, this means that I am inferior and powerless. If you do it out of pity, love, sympathy, generosity, etc., I am in debt, I have to be grateful. On the contrary, if it is my right I am not in debt: it is you that are indebted if you do not respect/satisfy my right. In general, different mental attitudes in compliance not only presuppose very different social relations, but make very different both the probability of the Goal Adoption, its readiness, and the future consequences for the social relations (for example in terms of credits and debts).

4. Normative Adhesion

As we said, Ns exploit and count on *a special process/kind of Goal-adoption*:

- First, they count on Goal-adhesion: that is, on the recognition by the addressee of the will of the issuer, and on an adoption due also to this: I adopt your goal also because I know that you want so.

[12] "Right" in this case means something like: "conform to a moral norm, to a value, to a law". Claiming a right is always searching for a shared value. See M. Miceli, C. Castelfranchi, *A Cognitive Approach to Values*, in "Journal for the Theory of Social Behaviour", Vol. 19, 1989, n. 2, pp. 169-193.

  - "Obedience" in general is a sub-kind of "Adhesion", and norm obedience is a kind of obedience.
- Second, it should ideally be motivated by the sense and respect of the authority and values; not by rewards.
- Third, it is a non "personal", individual request, but it is a generalized request, and should be understood as such and used as such.

4.1. Generalized Goal-adoption

There is an "individual" G-Adoption where
- X has to believe that Y (individual) has the goal that (DOES x A)
- and X comes to have (adopts) the Goal x (DOES x A)

There is a "generalized" G-Adoption where:
- X believes that there is a goal impinging not directly on a single individual but on a class or group of agents:
  (Bel x (Goal y (for any z member of C => (DOES z A))))
- if X believes to belong to that class,
- she believes to be concerned by the norm, and
- she instantiates a Goal impinging on her; adopts it.

However, having adopted the "generalized" goal X does not limits her mind and her behavior to this (self-regulation), she will worry about the others' behavior[13]:
- X is also able to have Goals about the others' behavior: she Adopts the Goal not to do but that for any z (DOES z A).
- Given such an Adoption she has *expectations* (predictions + prescriptions) about the others behavior, and is not only surprised, but "disappointed" by their non-conformity.

4.2. Spontaneous Norm Monitoring for Strong Reciprocity

A punisher has Adopted the goal that the bad guy behaves as prescribed and expected: she is not just "observing" but "inspecting" (surveillance).

She does not only have the mind of the norm-addressee (the "subject") but also the mind of the watcher, caretaker, and in a sense of the (re)issuer of

[13] Also because she is paying some cost for respecting the norm and the authority, for maintaining the prescribed social "order", which is supposed to be a "common". She wants the other be fair, reciprocate, contribute.

the prescription and norm[14]. This is why X also adopts the impinging goal of "punishing": this is not only a personal motive, an affective reaction, but it is also "expected" and prescribed, and approved by the others. And also this goal (to punish) is not only individual and personal, but is generalized:

X also expects that the others of the group would blame Z.

Actually, the famous expiation (penitence, amend) impulse in guilt feeling is the reflexive application of this goal to ourselves:

- the goal to be punished as *any* bad guy;
- and also the self-blame and reproach is already a self-inflicted punishment.

5. Against the Reduction of Norms to Sanctions, Incentives, and "Utility"

Obligations cannot/should not be hardwired, cutting the possibility of that behavior of that choice and agents must be assumed as responsible and free, and a true N presupposes the possibility of intentional violation.

Mental stuff is relevant. Obviously our behavior can be conformable to the norm but we are not "obeying" to it. Ns are not for accidental or mechanic conformity. An external "violation" is not enough for being "guilty" or "blameworthy".

Of course not only (and not always) "intentional" violation are guilty and to be blamed. What really matters is "responsibility"; that is, the counterfactual assumption that:

- "*X might have NOT done what he did*"; he had the possibility for behaving in a different way;
- "*X might have understood the consequences (harm, violation) of what was choosing*".

These, in fact, are the crucial assumptions (beliefs) supporting or eliminating "guilt".

We also reject the reductive "behaviorist" or "economic" views about "conformity" or "violation" just based on rewards, reinforcement learning and evocation, or on prediction of sanctions and rational decision for avoiding them. There are cognitive and social criticisms to that view.

[14] R. Conte, C. Castelfranchi, *Cognitive and Social Action*, cit.

Actually, sanctions are established and operative only for a sub-ideal world/case. Sanctions are in case of violation; but Norms does not expect/want violation!

Ideally N-Adoption is non-instrumental, not for convenience, just terminal.

The primary function of authority is not to monitor and to provide sanctions (even legal norms violations are weakly/rarely sanctioned (but symbolically they are)) is:

- to be recognized as the authority, to signal the existence of the authority and of the Ns;
- to issue the N as a N (that is "counting as" a N; recognizable as a N, not just a request or an abuse, etc.); is the "proclamation" the N, to be sure that it is common public knowledge and that it is "accepted" (and that there will be distributed social control: reissuing, confirming, monitoring, enforcing).

The second and secondary function of authority is to monitor (and to signal that it is monitoring), to sanction (and to signal that it will and is sanctioning).

The main function of prohibiting and of sanctioning (punishing) is signaling (the message: "this is bad!"), not the penalties (external costs): to stigmatize[15], to educate, to internalize norms and values[16].

In fact, no social control could compete with internal control[17], both, in surveillance (I hardly can hidden myself to myself), and in the certainty of the punishment.

We might have various kinds of normative minds/agents:

[15] S. BOWLES, H. SUNG-HA, *Social Preferences and Public Economics: Mechanism Design When Preferences Depend on Incentives*, in "Journal of Public Economics", Vol. 92, 2008, n. 8-9, pp. 1811-1820.

[16] G. ANDRIGHETTO, D. VILLATORO, *Beyond the Carrot and Stick Approach to Enforcement: An Agent-based Model*, in Kokinov B., Karmiloff-Smith A., Nersessian N.J. (eds.), "Proceedings of the European Conference on Cognitive Science", New Bulgarian University Press, 2011; D. VILLATORO, G. ANDRIGHETTO, R. CONTE, J. SABATER-MIR, *Dynamic Sanctioning for Robust and Cost-Efficient Norm Compliance*, in "Proceedings of the 22nd International Joint Conference on Artificial Intelligence" (Barcelona, 16-22 July 2011); G. ANDRIGHETTO, C. CASTELFRANCHI, *Norm Compliance: The Prescriptive Power of Normative Actions*, in "Paradigmi", forthcoming; F. GIARDINI, G. ANDRIGHETTO, R. CONTE, *A Cognitive Model of Punishment*, in Ohlsson S., Catrambone R. (eds.), in "Proceedings of the 32nd Annual Conference of the Cognitive Science Society", Austin, Cognitive Science Society, 2010, pp. 1282-1288.

[17] R.L. TRIVERS, *Social Evolution*, Menlo Park, Benjamin Cummings, 1985.

- Agents only sensible to legal & economic sanctions (Rational cheaters?)
- Agents (also) sensible to social approval or reputation (Social sanctions)
- Agents (also) sensible to internal rewards or to internal terminal values.

This is the kind of ideal agent Ns are addressed to and try to build. We have to characterize his mind.

6. "Internalization" (and Why It Matters)

Ns have to be internalized, this is a diffused claim. However ... what does this really means? Where is the model of this mental mechanism? Not only of "internalizing", but of doing something for an internalized N?

Does internalization mean a "value"[18], a "terminal goal"? Do Ns provide a "reason" for doing: "I should/I have to"; and why this is different from "I like" "I desire" "I want"...?

Or internalization is just a learned automatic rule? Or is it the calculation of possible sanctions?

Also because punishments, sanctions, rewards in general are not just "external", from the other agent observing us. Punishments are also endogenous and/or self-inflicted.

Moreover, these internal/intrinsic negative rewards

- can be of moral type: self-blame, regret; moral disgust; sense of indignity, lowering self-esteem, disapproval etc. (costs: rumination, etc.). Many (negative) emotions contain negative rewards, punishments.
- can be active "sanctions" that I provide to myself, although not necessarily intentionally (and consciously) but on the basis of an emotional reaction and activation (costs: damages, compensation, expiation,...).

Punishments & sanctions are not aimed just at trivial reinforcement learning, or at intimidating and inducing the agent at avoiding violations just in order to avoid sanctions (an economic reasoning). They are – in humans – mainly aimed at the introjection of a value, of a non-instrumental goal of obeying norms, of respecting the authority (message: "proclaiming" the norm)[19].

[18] M. Miceli, C. Castelfranchi, *op. cit.*

[19] G. Andrighetto, R. Conte, F. Giardini, *Le basi cognitive della contro-aggressione: vendetta, punizione e sanzione*, in "Sistemi Intelligenti", 2010, n. 3, pp. 521-532; G. Andrighetto, D. Villatoro, *Beyond the Carrot and Stick Approach to Enforcement:*

Ideally the N wants to be adopted for internal motives, not instrumentally to external incentives, not for external pressure and explanations, and for external control; but for and as "recognition" of the authority; and signaling such a recognition and subjection.

The paradox of human normative construction is that we use sanctions (punishments) in order to teach the other to obey to norms not for avoiding sanctions![20]

Only sub-ideally, only in case of violation (the norm has already be violated) we use sanctions. Only sub-ideally we decide to obey the norm just in order to avoid sanctions.

Why this is so important? Not only for the certainty of the monitoring (and eventually of the sanctioning) but because it favors the spreading of the N, not only based on examples but on teaching, shared values, rules, codes, ... And also because this kind of conformity internally driven is much less costly for society both in term of surveillance, violations, sanctions. Moreover, this internalization makes much more stable the recognized authority and the repeated internal "confirmation" of it and of the N.

6.1. Conformity and Punishments as Messages

Punishments and sanctions are mainly "messages"; they are not only aimed at materially and immediately harming you. They are aimed at communicating to you that:

- "We know that!", "We saw you!", "Do not believe that this is ignored, or hidden, or not noticed"
- "We blame this, and you!", "We want you know that we disapprove this as a fault, a defect, a violations; and that we consider you bad";
- "We want to sanction you; that you pay for this; to apply some penalty for this; at least a damage to your social image or reputation"
- "We want to punish you; that is that you learn from this experience, that in the future you avoid this, or you cannot do this again"
- "Your image is compromised; your reputation is in danger"[21].

An Agent-based Model, cit.; F. GIARDINI, G. ANDRIGHETTO, R. CONTE, *A Cognitive Model of Punishment*, cit.

[20] C. CASTELFRANCHI, *Emotional Support to Strong Reciprocity*, talk at the Workshop on "Moral Emotions", Roma, CNR, 2008, http://www.academia.edu/2040526/Emotional_Support_to_Strong_Reciprocity.

[21] G. ANDRIGHETTO, C. CASTELFRANCHI, *Norm Compliance: The Prescriptive Power of Normative Actions*, cit.

There are negative emotions just related to each of these meanings and situation: the feeling of be exposed to the other observation and judgment (embarrassment, worry, ...), the feeling of have been "discovered"; the feeling of being blamed; the feeling of a threat, of an incoming aggression; ...

Conformity too is a message (intentional or functional) not only about the attitude and intention of the subject, and his understanding and recognizing the N, being a good guy, etc. but also about the existence and validity of that N itself, its restatement and distributed re-issuing, the generalize expectation about its respect[22].

6.2. *Subjects Not Cooperators: The A-technical, Non-rational Nature of the Deontic "Ought"*

N claims that we adopt even (and it is even better) without sharing the "instrumental" nature of the N, without understanding/adopting its "function" or end.

My "plan" – as N subject, as adopter – is different from the authority's "plan". Consider as developmental example the mind of a mother pushing her child to brush his theet; and consider the higher goal of this goal into the two minds.

The mother wants her child to brush his teeth every evening, in order to avoid decay. The child adopts the goal (see Fig. 1) in order to obey his mother and to make her happy; he ignores and could not understand the real function of his behavior (the higher goals in the mother's mind). What, relative to the intentional behavior and the mind of the child, is just an external goal and a function, is an intended goal in the mother's mind.

Exactly the same kind of relation often holds between government and citizens[23]. Government pushes citizens to do something it considers necessary for the public utility, for some common interest; but it asks the citizens to do this for mere obedience or by using rewards or sanctions. It does not rely on the citizens' "cooperation", on their understanding of the ultimate functions of their behaviors, and on their motivation for public welfare; it relies on the citizens' motivation for obedience or for money or for avoiding punishment. We are not supposed to "cooperate" but to "obey" and execute!

[22] *Ibidem*.

[23] C. CASTELFRANCHI, *Scopi esterni*, in "Rassegna Italiana di Sociologia", Vol. 23, 1982, n. 3.

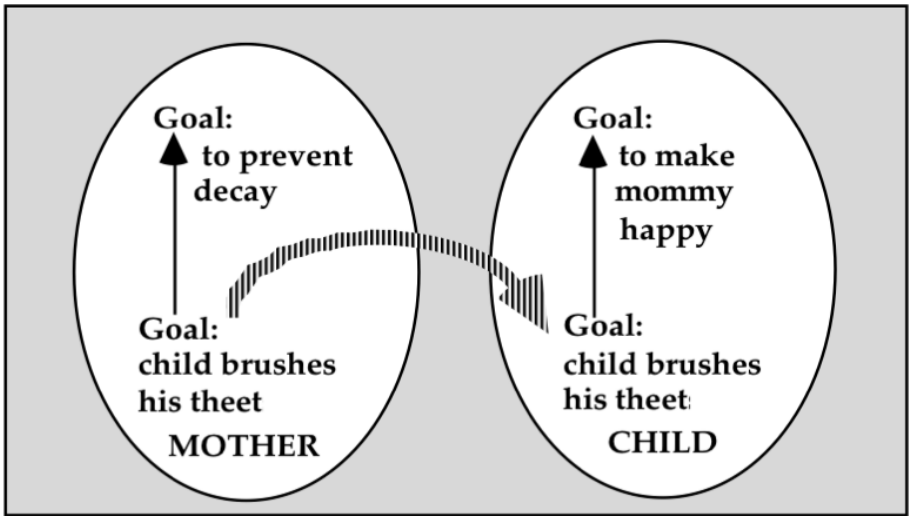

Fig. 1 – Goal-adoption from mind to mind

Both the "ideal" and the "sub-ideal" (for avoiding sanctions) obedience share a fundamental core, crucial for the real nature of the deontic "norm", "ought". A core that differentiate the mind of the normative "subject" S from the mind of the "issuer" or legislator.

S is not supposed to (have to) adopt the N (to "adhere" to the imperative) because s/he understands or agrees about its function, aim.

On the contrary, S is supposed to have to obey even if s/he does not understand the meaning of the N, or disagrees about it. This is true and mere "obedience".

A normative education is precisely an education to obey in any case, and even to not wondering and worry about the validity of the N.

In a sense the deontic "ought" "have to" is a de-technicalized "ought": no longer a necessary means for …something that you have to want, to choose[24].

The technical ought is: "If you like/want to/in order to …, you should, have to…" for example: "To open this door you have to…".

The deontic ought just is: "You have to" for what? why?

[24] I am working on this topic with Luca Tummolini (in preparation).

In the mind of the "issuer" the N is on the contrary supposed to be a means, a solution for some problem; he should have in mind the aim and instrumental function of the N.

6.3. Educating to Norms

There are two fundamental kinds of N education:

One based on Understanding & Empowerment/Responsibility: "Did you see what happens if/when …"; "Do you understand what you have done?"; "You must never do that, otherwise there will be this trouble".

And the most radical and "duty" based one: No understanding/explanation, no sharing.

A radical "normative education" starts when your mother moves from just saying: "*Don't say dirty words!*" "*I don't want that you say dirty words!*" or "*You should not say dirty words!*" to an "impersonal" formulation: "*One should not say dirty words!*" "*Dirty words should not be said!*" "*It is bad …!*". And when to your protest or question "*Why*?" she does not give any explanation, but just answer: "*Because otherwise I bit you!*" or "*Because I want so!*"; or even better something like "*It is not allowed; and that's all!*" "*You must obey; that's it*" "*Because it is so!*".

That is, she refuses to give you justifications and reasons, and teaches to you that you should do this without specific instrumental reasons (and advantages), terminally; just because it is an order, a norm, of an authority which should be acknowledged; as a terminal "value".

6.4. The "Alienated" Nature of Norm Adoption

Contrary to what supported by the theory of the "extinction of the State" and of the government not over persons but of persons/people (as conscious cooperators and intelligent planners of social dynamics), in my view, it is not really possible the total elimination of this alienated cooperative relation: to cooperate to common good (by conforming to it), without understanding it and thus without intending it.

It is not possible a citizen fully aware and intentionally cooperating with public choices, to whom nothing is thus "imposed" - including imposts. At least partially he will remain a "subject", not sharing the ends but just obedient.

It is not possible for cognitive and cultural limitations of individuals and groups[25]. It is in any case unavoidable a portion of "delegation" and not understanding of the reasons and aims of public/collective choices. Ns and the "authority" of the authority are for that.

Possibly this "delegation" should be based on "trust" towards rulers and institutions, towards the dominating groups, not on the awe of their power and of possible harms.

That the "subject" of the N could not/should not understand (share) the technical sense of the N (that this be irrelevant for her/his obedience and that her/his technical evaluation be irrelevant for the collective) for sure is not a warranty that the N is not oriented at aims different from the "common good".

I, the subject, trust them or the system or anyhow undergo to them, I cannot understand, dominate those issues. Will the N really serve more to the ends of a class or part of the society? Will it protect the interests of the legislators?

There is "alienation"; meaning that the subject alienates his own intellectual capabilities of evaluation, problem-solving, decision, by "delegating" to others them, and the power and the solution. Moreover, he is not in condition to realize that, to understand this process, and behaves without recognizing his own estranged powers and without the possibility of reappropriating them. Perhaps Y – delegated to that – has really found the right solution for our collective problem, but I am not supposed/required to understand it, share, adopt as such. I have to adopt it blindfold.

7. Influencing Devices in a "Prevention Focus"

Norms are influencing devices. In principle they might

- either use a "positive", promoting, promise-based perspective: an attraction perspective,
- or use a duty-based one: ought, responsibility, guilt, punishment: an aversive perspective ("Prevention focus")[26].

[25] For problems of timeliness and regularity in the coordination of distributed and countless behaviors; for the spontaneous psychological prevalence of personal, group, local, and short term interests, on the collective and long term ones, with their necessary compromises and sacrifices.

[26] E.T. Higgins, *Promotion and Prevention: Regulatory Focus As a Motivational Principle*, in Zanna M.P. (ed.), "Advances in Experimental Social Psychology", San Diego, Academic Press, Vol. 30, 1998, pp. 1-46.

Why they generally adopt the second one, and build "obligations", "constraints", "duties", . . . ?

The logical formulation seem to be equivalent, interchangeable; but they are not at all equivalent from the psychological (cognitive, motivational, affective) point of view[27].

Usually laws and norms do not attract and promise, but present the threat face (on 10 Commandments 8 are prohibitions).

Also because – as we said – they do not want to explain the (instrumental/technical) advantage of the norm, its "utility" (they just want that you want to be "conform", to obey).

There is an asymmetry in the use of threats and promises, sanctions and rewards, blame and bad reputation vs. praises. Why? Why the main tool and artifact for social influencing and social order control adopts such a avoidant/threatening strategy? Betting on the "negative" side ("Prevention focus") rather than on the "positive" one?

(i) "Surveillance"

One reason is precisely the control and in particular the "monitoring". It is necessary to consider the possibility of "violation"; and must be focused and alerted on it, in order to prevent or to sanction it.

(ii) "Punishment" costs and incomes

One of the reasons/functions of this shaping and asymmetry is that it is preferable to support an altruistic cost in order to punish a deviant, a violator, more than wasting resources for praising, reinforcing a good conforming guy. In fact, it is not simply a matter/aim of education and of educational means (where the positive rewarding approach might even been more effective), but it also is a problem of compensation, balancing: who has violated have to "pay" for that, and a sanction (even just blame, emargination, bad reputation) is an harm, a paid penalty.

(iii) A more effective mental frame

Another reasons is due to the supposed greater efficacy or greater weight of the punitive and dutiful approach. It bets on losses, that – with equal amount – worth more than earnings; losses, risks, possible harms have a psychological priority, and more value in evaluation and decision making ("Prospect" theory); they are more binding.

(iv) A "natural" link: punishing for inducing avoidance

[27] C. CASTELFRANCHI, M. GUERINI, *Is It a Promise or a Threat?*, in "Pragmatics & Cognition", Vol. 15, 2007, n. 2, pp. 277-311.

There is a logical and affective direct relation between a "punishing" experience or threat (which induce an avoidance reaction and feeling) and the aim to build an avoidance attitude towards the violation of the N.

"Duty" is an "avoidance goal"; is a constraining representation.

8. Norm Processing from Beliefs to Goals and Intentions

The ideal path of N in our mind is:

- x is able to *recognize* N, to differentiate what is a N and what is not;
- x is able to *assess* whether s/he is concerned by N;
- x *accepts* N, forms a N-goal corresponding to N;
- x *decides to comply* with N or not;
- x is able to *re-issue* N, to prescribe it to other fellows subject to N, and
- x is able to observe, *monitor* their behaviors with respect to N and react in a positive or negative way to them[28].

In fact, for a N to influence x's behavior, N must become an *intention* of x.

For something to become an intention, it must first become a *goal*, which will be decided upon, planned etc. by x. Finally, for a new goal to be formed, x must form:

- beliefs about its reasons;
- the normative goal (norm acceptance).

Eventually, decide to achieve it (*norm compliance*) (Fig. 2).

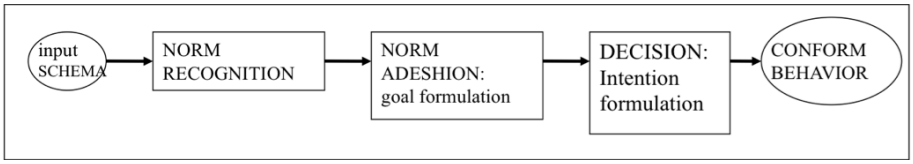

Fig. 2 – Cognitive norm processing

More precisely:

- beliefs: the recognition of that practice or request as a "norm"; and as impinging also on me; the acknowledgment of the "authority";

[28] R. Conte, C. Castelfranchi, F. Dignum, *Autonomous Norm Acceptance*, in Mueller J. (ed.), "Proceedings of the 5th International workshop on Agent Theories Architectures and Languages" (Paris, 4-7 July 1999).

- beliefs: the instantiation of that N in the current situation and context;
- goal Adoption: (N Adhesion) the formulation of the normative goal potentially regulating my action in the situation;
- the activation/consideration of possible Motives (higher-goals) for obeying/conforming or not;
- decision to violate or to obey: possible formulation of the normative Intention;
- intention in action and conform external behavior;
- success or failure (possible non intended violation due to an accident).

Along this path there are several reasons for dropping a N-goal:

- goal-conflict: the N-goal contrasts with another goal of the agent:
  - probability and weight of punishment
  - importance of the goal or value of respecting the norm
  - importance of feelings associated to N-violation
  - importance of the negative consequences of violation
- N-conflict: N contrasts with other Ns accepted by the agent
- irrelevance: x does not believe to be a member of the set X
- material impossibility: x forms a N-goal but cannot comply with it.

9. From Goal-adoption, Decision, Intention, ... to Routines

Our quite rich cognitive characterization of the representations and processes underlying a behavior obedient to a norm, should not however give the idea of behavioral conformity as always based on such a complex "reasoning" and "deliberation".

It is absolutely true that norm conformity and obedience can become a habit, an automatism, a routine behavior, based on simple production-rules or "classifiers".

By default – except one has special reasons and active goals blocking the trivial reaction and routine – one just executes the classifier (like when - while driving a car – one reactively stops to a red light):

Condition ==> Action;

Recognized stimuli ==> Appropriate behavior.

Given that normative behavior is a "regularity" (norms implement and maintain regular and common behaviors), there is a regularity both in perceiving (a fixed schema) and in acting (a fixed behavior in those conditions); thus, reasoning and decision become superfluous (wasting time and resources).

Normative routine behavior, in our model, is just a "shortcut", a functional bypass of the original and "normal/ideal" way, which is assumed to usually be its origin and source, and its cognitive background and justification (Fig. 3).

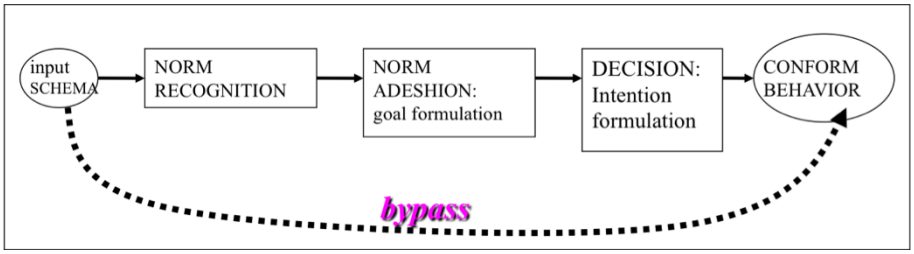

Fig. 3 – Normative routine behavior

This is the simplified schema of reflex "obedience": like stopping at the red semaphore.

However, there are top-down vs. bottom-up processes, moving from a initial mere reinforcement or imitation based learning to a real deontic representation and awareness, or vice versa, moving from a deep normative understanding and processing to a simplified routine and reaction.

When respecting certain norms is fully routinized and becomes a mere automatic unconscious reflex, the subject is no longer aware of the norm, that is, he does not activate the normative belief ("It is prohibited . . . ") and the decision.

However, on the one side, even for fully automatic conform behaviors it is always possible the evocation and explicit consideration of the normative beliefs and goals. For example, we automatically stop at the red light, but an ambulance with a loud siren arrives behind us and "asks" us to move, to cross; in that very moment we might explicitly consider that it is red and it is prohibited to cross, but also that it is better to violate that norm.

On the other side, for several norms that level of unconscious automatic obedience is not possible, and we consider the existence of the prescription and take real decisions. For example, stopping for a policeman' signal is not fully automatic; paying a fine or taxes requires some reflection and a conscious deontic reasoning; and so on.

Moreover, what matters for social order, is that it is "as if" the agent was remembering and considering that there is a norm; we implicitly rely on that.

10. Norms As Multi-agent Artifacts

Ns are social and mental artifacts. As "social" they are a "coordination artifacts", but this coordination works by coordinating agents' minds and mental attitudes. Normative coordination presupposes in fact different normative roles with their minds.

Normative minds: the "Issuer" I; the "Subject" S; the "Monitoring agent" or "Watcher" W.

Those attitudes are complementary to each other, and necessary for the social implementation of Ns:

- In all those "minds" the N is an imperative on a class of agents.
- However, in I and W's minds the N does not concern them; it is not "instantiated" on them.
- S on the contrary is concerned, and should arrive to formulate a conform Intention to do.
- I and W instantiate the N on S, and formulate the Goal (Expectation not just forecast) that S behaves correctly.

We have – in this paper – gone deeply into S's mind and his adoptive process; this was the cognition and internalization we have focused on; but we have also explained how S becomes a (non official, informal) W and I.

11. Concluding Remarks

In sum, to work appropriately Norms must be mentalized. Norms are for influencing "autonomous" agents, that is, agents self-regulated and self-motivated; for inducing goals in them.

Norm working cannot be reduced to sanction/incentives and "utility". Ideally N-Adoption is terminal, non-instrumental; they should not be obeyed because of possible "sanctions".

Norm must be recognized and acknowledged as such; they cannot be only implemented in routines and habits. The N spreading and maintenance is first of all a mental spreading and sharing of values and beliefs.

Norms are a behavioral and mental coordination artifacts, based on different complementary mental attitudes in the various roles.

Expectations about the others' behavior conform to a norm are not just "predictions" built on a regularity; they are full expectations, entailing the

fact the we rely on the other behavior and thus want/wish it; we not only predict but prescribe the others' behavior[29].

There is a specific "cognitive processing" of norms in cognitive agents, from the recognition of the input as a normative prescription to the formulation of the intention to conform or to violate.

[29] C. CASTELFRANCHI, L. TUMMOLINI, *Positive and Negative Expectations and the Deontic Nature of Social Conventions*, in "Proceedings of the 9th International Conference on Artificial Intelligence and Law - ICAIL 2003", ACM, 2003.