The *Normative* Framework of

COOPERATION

&

the *Cooperative* Nature of Social Norms

Cristiano Castelfranchi & Luca Tummolini
GOAL group

PROJECT *SOCCOP* - TECT

***CNR- Institute for Cognitive Sciences and Technologies* -Roma**

Not a complete and systematic analysis/theory of the cognitive foundation of cooperation

&

of norms (conventions, social, moral, legal, ...)

as a fundamental solution for cooperation problems ;

just some *basic issues* and *lines*.

Cooperation among humans is not just “behavioral”

Before being behavioral must usually be “mental”:

due to psychological representations, attitudes, decisions that teleologically guide behaviors.

- **Which are the mental representations and processes that mediate human cooperation?**

We will examine only some (crucial) aspects of that problem:
goal-adoption; goal-adhesion; expectations and prescriptions
about the other mind and behavior.

In particular:

-the prescriptive, deontic, *normative nature* *that human coordination and cooperation* acquire;

- the *adoptive/cooperative nature of norms* and their 'obedience';

- the nature of (social) *norms* as a *cooperative coordination artifact*.

1. **Our perspective: *Representations* and *Goals***
2. **C*ooperative Cognition*: Goal-Adoption theory**
3. ***Normative Cognition*: Some crucial issues**
4. **Normative "Adoption"**
5. **Internalization**
6. **Conclusions**

1

Our general PERSPECTIVE

The *Centrality of* ***GOALS***

- **"Cognitive" does not means "epistemic"**
- ***Mental representations*** *=/= knowledge, beliefs, ...*

It also means "Goals": ***mental representations*** *about what should be(come) true.*

- ***Mind reading*** is not only about understanding, predicting, coordinating, sharing, ..; **it is for having Goals about your mind, for changing your behavior by *manipulating* your mind**.

Our general PERSPECTIVE

The *Centrality of* ***GOALS***

• In our view, Economics, GTh, Logics, Primatology, etc. do not put enough attention on Goals, and on their social dynamics: ***Goal-Adoption, Goal-Delegation, Goal-Induction***, ..

Beliefs are just for managing Goals, since only Goals control our behavior.

In particular:

- **Goal-Adoption/ Goal-Adhesion**

Also Norms are for ***influencing*** our behavior by *changing our mind*, our preferences and intentions (by changing our beliefs):

> they are **goals about our goals** (that 'control' our behavior).

2

‘Cooperative’ Cognition & ‘Normative’ Cognition

‘Cooperative’

Cognition

&

‘Normative’

Cognition

Goal-Adoption

How the Mind becomes “social”

Goal-Adoption

How the Mind becomes "social"

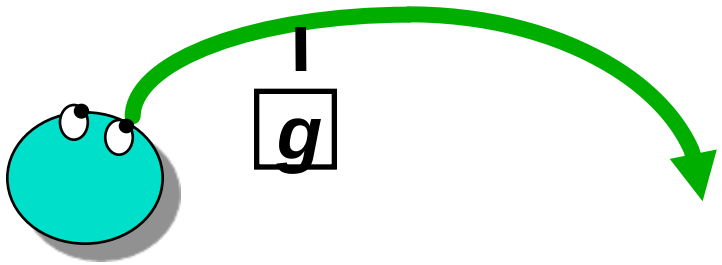
>> **Mind reading is for**

'goal-adoption' and ***'goal sharing'***

>> **Mind reading is for**

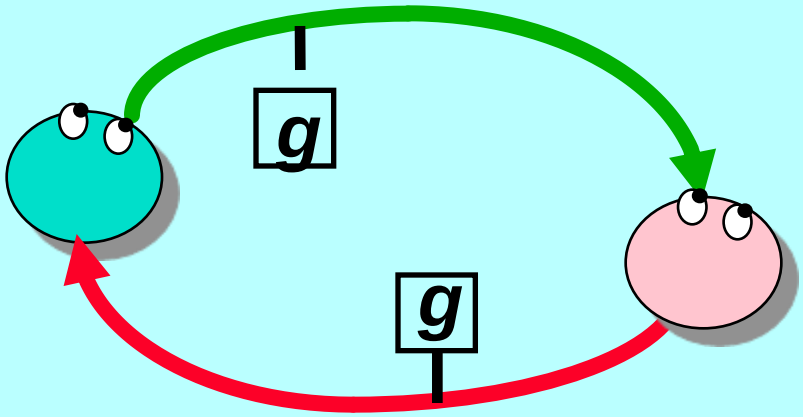
'goal-induction' and ***'manipulation'***

SOCIAL AGENTS - Micro-Sociality



Delegation **or** ***Reliance***
to exploit

Goal adoption
to help



> DELEGATION / RELIANCE:

X realizes her GOAL
thanks to Y's powers and action

> Goal-ADOPTION :

Y adopts and pursues X's GOAL

NO real/effective COLLABORATION without

UNDERSTANDING

&

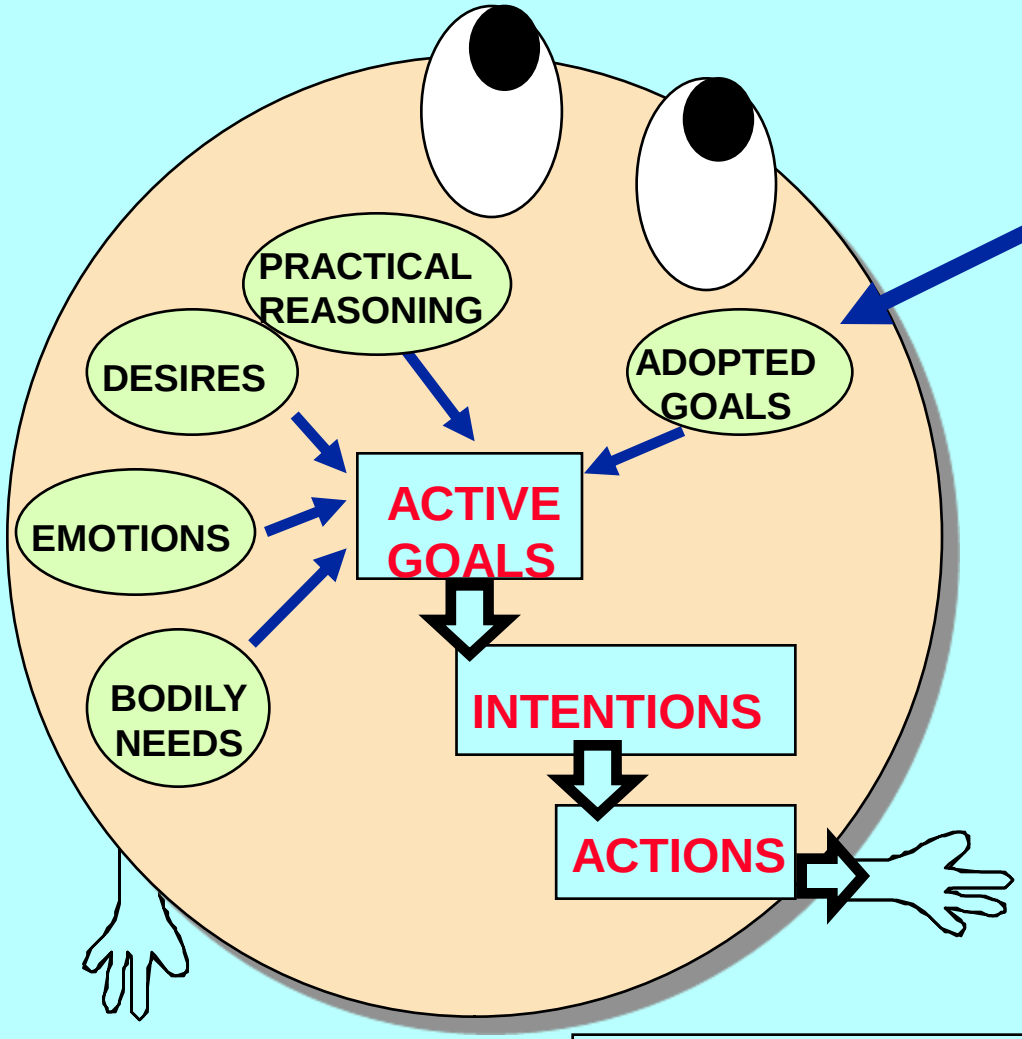
ADOPTING or COUNTING-ON

the GOAL of the other

- **You have *goals* about my mind;**
- **I have *goals* about your mind, not just *beliefs*:**

I give goals to you and receive goals from you.

Social-Agent's Architecture and Multiple Goal-Sources



Help,

Requests

Promises

Norms,

……..

beyond strict BDI

the ‘prodigy’ is that those *self-regulated*, goal-driven systems can *import goals* from other goal-driven, purposive systems, from outside:

- They put their ‘body’, skills, problem-solving capacity, and resources at disposal of the needs/desires of another agent.
- They spend their powers, and actively pursue the goal of another and for another; and , vice versa, **Y exploits X’s body for her purposes**.

“Auto-nomos”,“self-motivated”,“Goal-driven” …

doesn’t mean: **“Selfish”**

Goal-ADOPTION

"X has the Goal G1 since and until it is the Goal of Y"

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.

This is 'goal-adoption', and can be ***motivated*** by different reasons.

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

Goal-Adoption is not 'IMITATION'

is not "doing the same", "doing like the other"

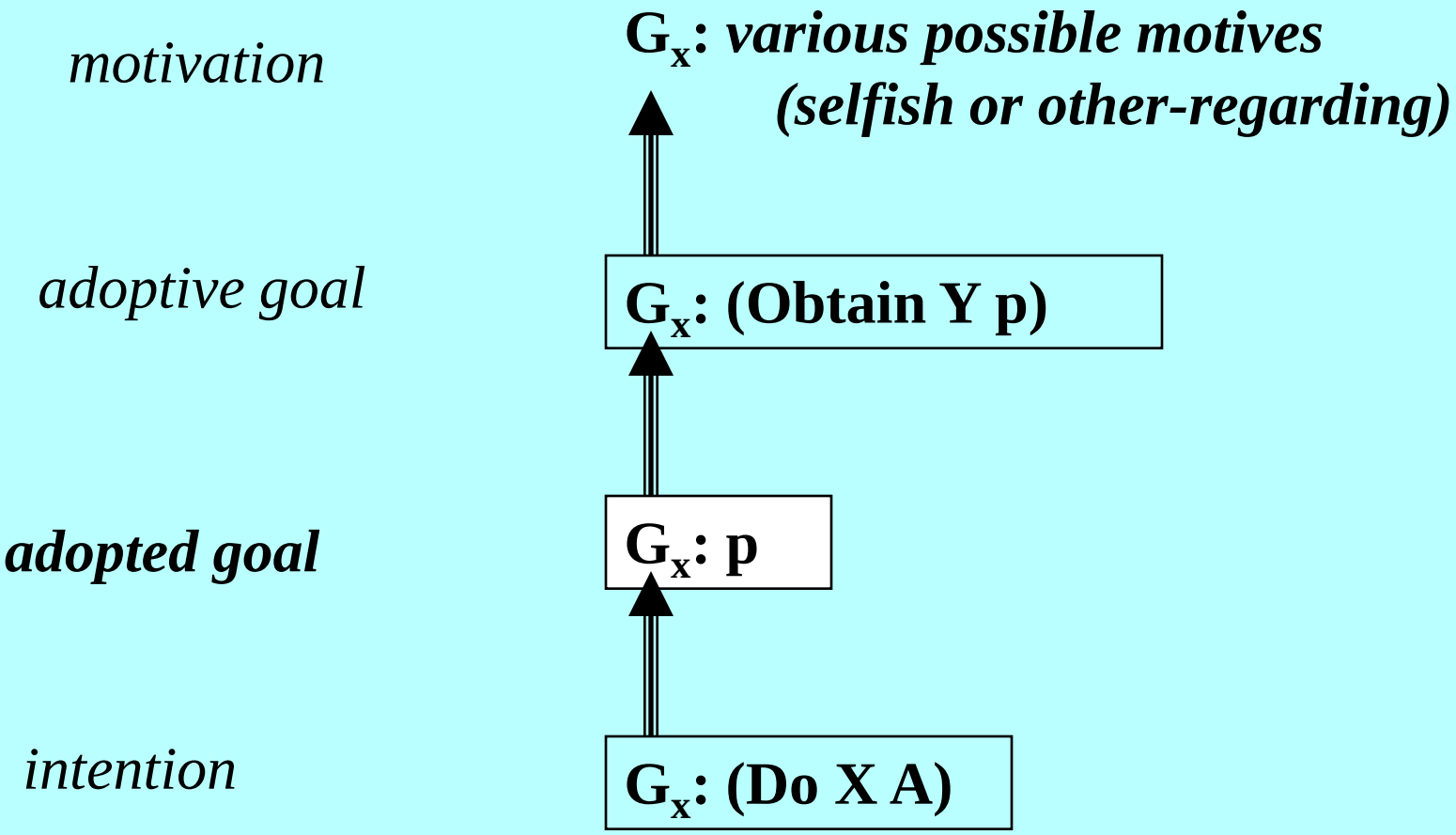
It is doing something 'for' the other,
for realizing her Goal

[Imitation is not enough for Cooperation:
***complementary* and *substitutive* activities]**

Goal-Adoption is not 'ALTRUISM'

It is doing something 'for' the other,
for many possible motives:
including selfish advantages; like in exchange and commerce

A Complex Goal Structure



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.

>> **a double level of adoption** (a meta-adoption): *I know and adopt your goal that I adopt.*

Moreover, in case of Adhesion there is an agreement between X and Y about X's adoption, X doing something as desired by Y.

3

‘Cooperative’

Cognition

&

‘Normative’

Cognition

Two crucial Issues

- **Not just "constraints"**
- **From "Beliefs" to "Goals"**

Not only "constraints"

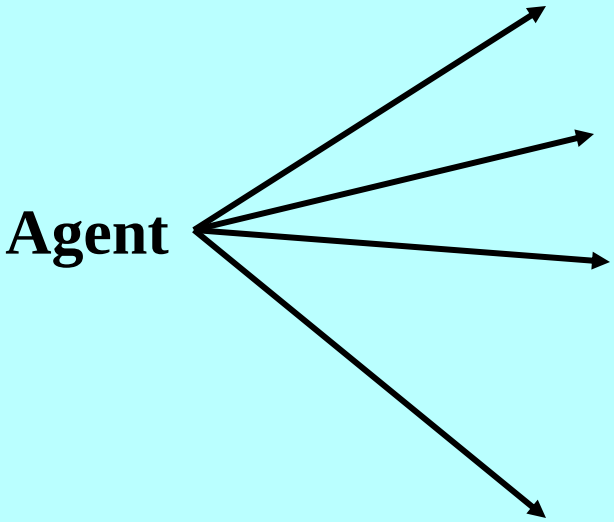
In many organizational, anthropological, sociological views not only there is a very strong (if not exclusive) **emphasis on "sanctions" as necessary and "definitional" for having "norms",**

> but there is an explicit or implicit view of N (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 N 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.

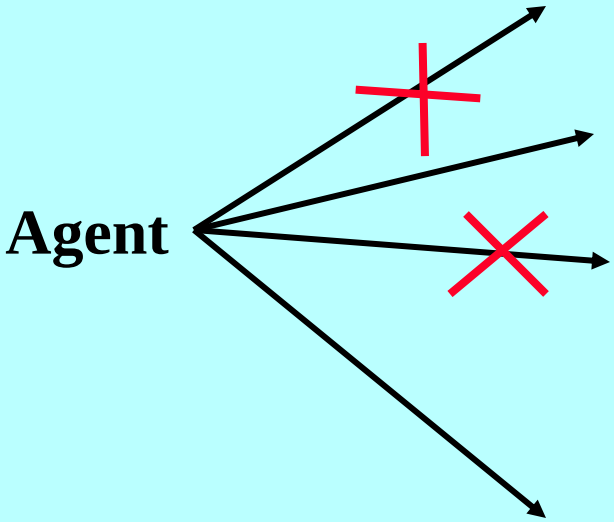
Not just ***blocking*** some possible choice or changing the evaluation by altering the expected outcomes (rewards) of the alternatives.

Not just **“pruning”** possible actions, or “permitting” them:



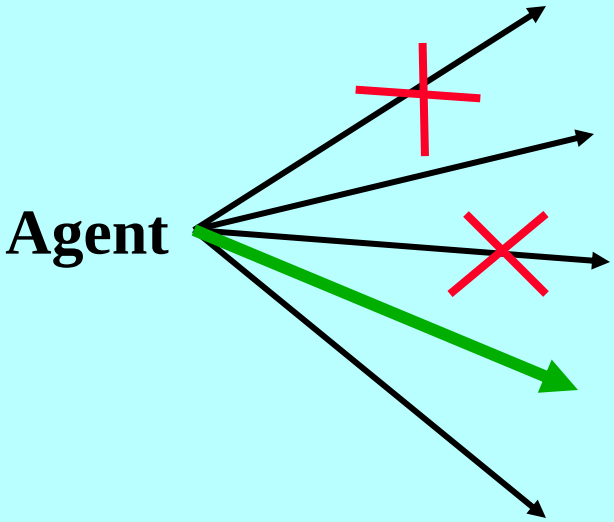
Not just ***blocking*** some possible choice or changing the evaluation by altering the expected outcomes (rewards) of the alternatives.

Not just **“pruning”** possible actions, or “permitting” them:



Not just ***blocking*** some possible choice or changing the evaluation by altering the expected outcomes (rewards) of the alternatives.

but ***adding***, <u>creating new goals and alternatives:</u>



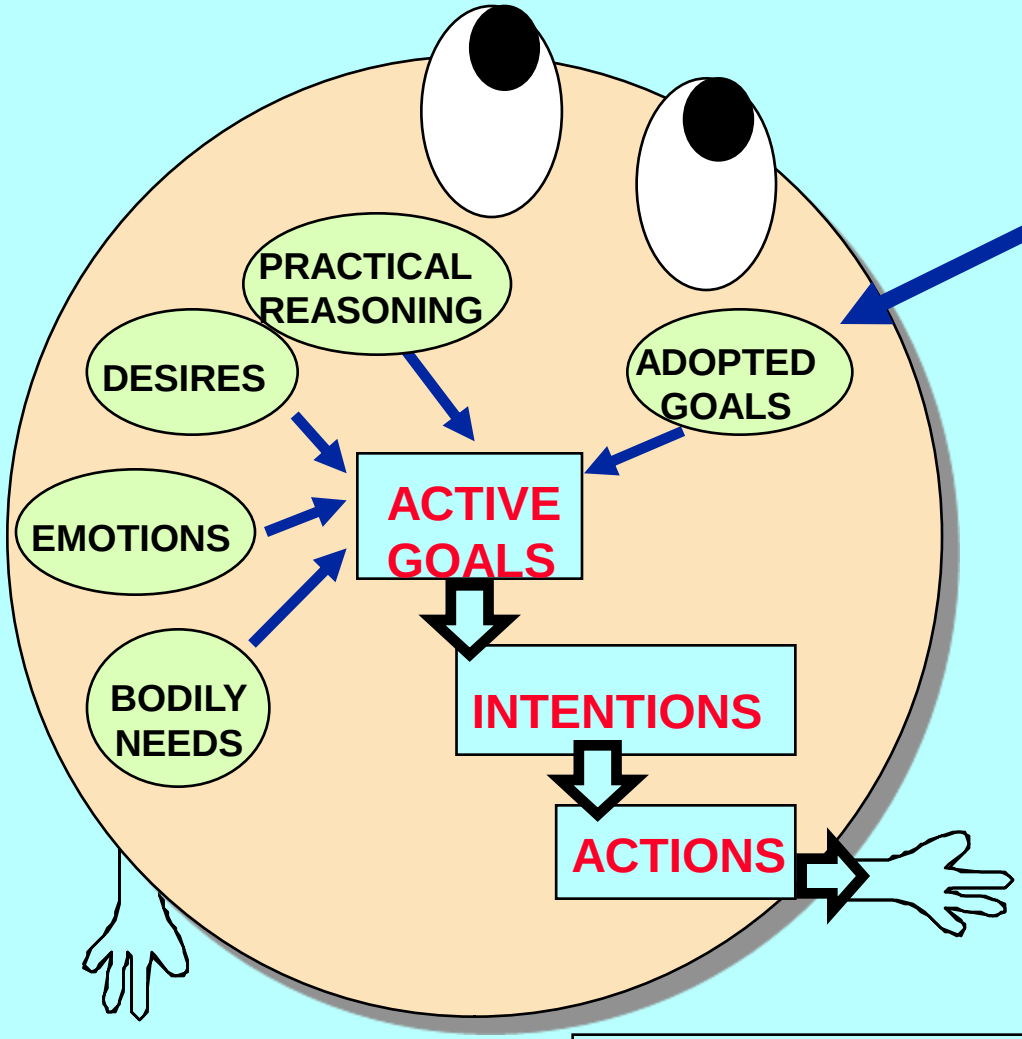
Not only "constraints"

In this view, it seems (it is implicitly assumed) that: **goals (and then intentions) of the agents 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***), they are imported, "adopted"; they are "prescriptions" and "imperatives" from another Agent (the authority).

Social-Agent's Architecture and Multiple Goal-Sources



Help,

Requests

Promises

Norms,

……..

beyond strict BDI

Not only "***constraints***"

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.

"COOPERATION" is also based on *influencing* (promises, threats, agreements,..), *duties*, *debts*, *obligations*, etc.

From *BELIEFS* (about the others' behavior and mind)

to *GOALS* (about the others' behavior and mind)

An ideal path:

1. From '*Predictions*' to '*Expectations*'

I have not only a *Belief* but also a *Goal* about the other's behavior: I "wish", want; and I hope, trust, that the other will act in the expected way.

But it doesn't depends on me (in my mind: Belief).

The hybrid nature of expectations

Expectations are not mere *beliefs* (“*predictions*”).

Expectations imply a **subjective concern about the realization of the anticipated event**

>> I’m “expecting” “waiting for”… something

>> There is some *goal* involved

That’s why they can be “positive” or “negative”

Positive Expectation **Negative Expectation**

Bel (x p^{fut})	***Bel (x p^{fut})***
Goal (x p^{fut})	***Goal (x Not p^{fut})***

Bel$^{\%}$ = Beliefs have *degrees of certainty*: **one can be *pretty sure that, enough sure, 50%*, etc.**

Goal^{v} = Goals have *value*, they can be *more or less important*

We assume that

the intensity of the emotional* reactions *related to an Exp (surprise, disappointment, relief, ..)

is function of its components (cognitive ingredients)

From *BELIEFS* (about the others' behavior and mind)

to *GOALS* (about the others' behavior and mind)

An ideal path:

2. From "Expectations" to "Prescriptions"

<u>Beliefs</u>: it also depends on me; I can *influence* their behavior

From the *<u>Goal</u>* that they act in the expected way to the *Goal of "influencing"* them to do so:

From *BELIEFS* (about the others' behavior and mind)

to *GOALS* (about the others' behavior and mind)

An ideal path:

2. From "Expectations" to "Prescriptions"

> first, I have the Goal of influencing them at least by letting them know that I'm relying on them, counting on their right behavior (they will take into account this; by mere 'adjusting' or by 'adopting' my goal);

> then, because they let me count on this, they allow and entitle me to ; they get a commitment, an obligation;

> I have the Goal that they *adhere* because I'm entitled and they recognize so: a prescription (in a broad sense).

From *BELIEFS* (about the others' behavior and mind)

to *GOALS* (about the others' behavior and mind)

An ideal path:

3. From "Conventions" to "Norms"

The transition from mere "conventions" to real social/moral "norms" is along this path:

Conventions can start as merely **"predictive"** (Beliefs-based); Norms finally are **"prescriptive";** not only based on *Goals* and Adoption, but based on *Influence* and Adhesion.

The reason why Norms are not only Predictive but Prescriptive (**influencing devices**) is that not only I rely on your behavior and want you act in the expected way, but there might be some temptation, some selfish advantage for you in violating the N. The N wants influence you to do as prescribed, since you might have preferences for violating (Gintis).

I have ***beliefs*** about the others behaviors and their predictions (beliefs) about my behavior (Bicchieri), but:

- I have also beliefs about the fact that they rely/count on my right behavior (Bicchieri); that is, they also **want/wish** so (Goal)

(why they should be upset and aggress me, *punish* me, if they wouldn't ***want*** that I do in the other way?)

- I have beliefs about the fact that they know that I know about their expectations and reliance, .. and I let them expect so and count on this; thus I ***entitle*** them to rely on this; they will feel entitled; I'm tacitly committed to the right behavior;

(why they should *sanction* me? If they wouldn't feel that I'm violating, betraying, ...)

>> **They not only have "*reasonable* expectations"** (Lewis), **they have "*entitled* expectations" (by me). ("Tacit agreement")**

Spontaneous, self-organizing human *coordination*

spontaneously acquires a normative nature,

and

this *cooperative solution* works also thanks to

the psychology of 'obligation', 'commitment', 'violation', …

4

‘Normative’

Adoption

Norms 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.

"Obedience" in general is a sub-kind of "Adhesion", and norm obedience is a kind of obedience.

• Second, it is a non "personal", individual request, but ***it is a generalized request, and should be understood as such and used as such***.

•Third, it should ideally be ***motivated by the sense and respect of the authority and values; not by external rewards***.

Norms exploit and count on

a special process/kind of Goal-Adoption

• First, ***they counts on Goal-Adhesion***: that is on the recognition by the addresse 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.

"Obedience" in general is a sub-kind of "Adhesion", and norm obedience is a kind of obedience.

• Second, it is a non "personal", individual request, but ***it is a generalized request, and should be understood as such and used as such***.

•Third, it should be ***motivated by the sense and respect of the authority and values; not by external rewards***.

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

Generalized Goal-Adoption

but, having adopted the 'generalized' goal

X doesn't limits her mind and her behavior to this (self-regulation), ***she worries about the others' behavior***:

- 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 and prescriptions) about the others behavior**, and is not only surprised, but 'disappointed' by their non-conformity.

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 doesn't 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 the prescription and norm** (Conte & Castelfranchi, 1995).

NORMS

as

MULTI-AGENT artifacts

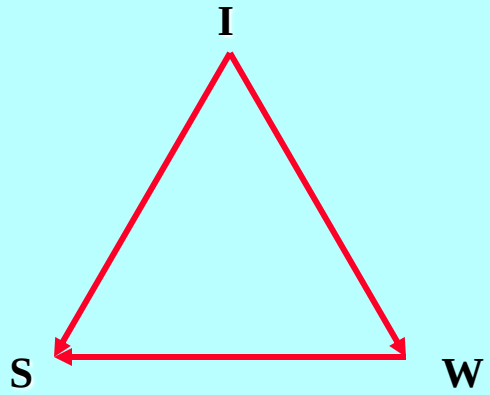
NORMS as *MULTI-AGENT* artifacts

N as a Multi-Agent Notion and Object

Normative mindS: the mental ***attitudes*** of different necessary normative **Roles** (not necessarily by different individuals):

the "Issuer" I; the "Subject" S; the "Monitoring agent" or "Watcher" W

Those attitudes are *complementary* to each other, and necessary for the norm social implementation:



The efficacy of normative regulation implies a "*cooperative*" attitude:

---

1. **Norms work thanks to the *cooperation and compliance of the agent:***
    - ***while recognizing the N as a N;***
    - ***acknowledging the authority;***
    - ***adhering to the request/prescription;***
    - ***deciding to conform;***
    - ***monitoring the others' behavior;***
    - ***blaming and sanctioning;***
    - ***educating; etc……***

The efficacy of normative regulation implies a "*cooperative*" attitude:

---

2. Norms work thanks to the *cooperation among different roles and subjects*, and the complementariety of their mental attitudes and consequent behaviors

Norms exploit and count on

a special process/kind of Goal-Adoption

• First, ***they counts on Goal-Adhesion***: that is on the recognition by the addresse 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.

"Obedience" in general is a sub-kind of "Adhesion", and norm obedience is a kind of obedience.

• Second, it is a non "personal", individual request, but ***it is a generalized request, and should be understood as such and used as such***.

•Third, it should be ***motivated by the sense and respect of the authority and values; not by external rewards***.

Norms exploit and count on

a special process/kind of Goal-Adoption

Like 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. (Castelfr. AI&LAW, 2001)

So, the N wants:

(i) my *behavior*, due to

(ii) my *goal*, due to

(iii) my *adhesion*, due to,

(iv) *motivated by* (higher-goal) internalized non-instrumental values.

Norms exploit and count on

a special process/kind of Goal-Adoption

• A real ***"normative education"*** starts when your mother passes 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 doesn't just answer: *"Because otherwise I bit you!"* or *"Because I want so!";* but 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, a terminal 'value'.**

AGAINST THE REDUCTION OF NORMS TO SANCTIONS, INCENTIVES AND "UTILITY"

Cognitive and social criticisms

- **Sanctions are only for a *sub-ideal world***
- **Sanctions are in case of violation, but Norms doesn't want violation; doesn't expect to be respected for violation.**

Ideally N-Adoption is terminal, non-instrumental, for convenience

The main function of prohibiting and of sanctioning (punishing) is *signaling* (the message: *"this is bad!"*), not the penalties (external costs):

- **to *stigmatize*** (Bowles & Hwang), **to *educate*,**

  to *internalize* norms and values.

‘Internalization’ ?

No *social control* could compete with *internal control* (ex. Guilt feelings) (Trivers), **:**

Both, in ***surveillance***

(I hardly can hidden myself to myself),

and in the ***certainty of the punishment*.**

Internalization (and why it matters)

Punishments & sanctions are not aimed just at a 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** (Miceli & Cast), of a non-instrumental goal of obeying norms, of respecting the authority (*message*: **"proclaiming"** the norm)

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*! (Castelfr).

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

"Internalization"??????

We agree about the need for INTERNALIZATION, etc. however.....

What does this really means?

Where is the model of this mental mechanism?: not only to "internalize", but to do something FOR an internalized N?

>> Does this mean a "Value" (Miceli**), a "terminal goal"**

Norms provide a "reason" for doing: "I SHOULD/ HAVE TO"; why this is not "I like" "I desire" "I want"..??

or

>> Just a *learned* automatic *rule* ?

Concluding Remarks

Cooperation presupposes and exploits specific mental attitudes and processes:

- ***Goal-adoption;***
- ***Goal-delegation, influencing & then relying (trust)***

both require *mind-reading*

- ***Expectations* and then entitled *prescriptions* on the others behaviors;**
- **the emergence of *social conventions and norms* based on *tacit agreements***
- ***Norms* as influencing and cooperative devices**

- **NORMS ARE FOR INFLUENCING "AUTONOMOUS" AGENTS, that is, AGENTS SELF-REGULATED AND SELF-MOTIVATED; by <u>INDUCING GOALS</u> in them**

- **AGAINST REDUCTION OF NORMS TO *SANCTION/INCENTIVES* AND "*UTILITY*"**

- **AGAINST REDUCTION OF NORMS TO *ROUTINES***

- **AGAINST REDUCTION OF *EXPECTATIONS* to *BELIEFS***

Not a complete and systematic analysis/theory of the cognitive foundation of cooperation

&

of norms (conventions, social, moral, legal, ...)

as a fundamental solution for cooperation problems ;

just some *basic issues* and *lines*.

We like to thank TECT Project, a really advanced and interdisciplinary joint research

We like to thank our research group in Cognitive Science at ISTC:

A part from *Rosaria Conte (LABSS Group)*

Maria Miceli
Rino Falcone
Emiliano Lorini
Fabio Paglieri
Giovanni Pezzulo
Michele Piunti

END

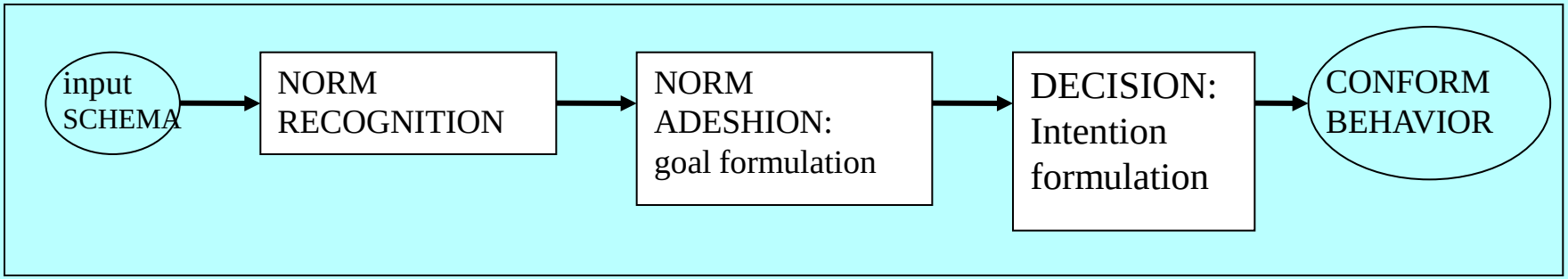
5

6

From Goal-Adoption, decision, intention, … to ROUTINES

Norm conformity as *routine* behavior

Our quite rich cognitive characterization of the representations and processes underlying a behavior obedient to a norm, ……



……. shouldn't however give the idea of ***behavioral conformity as always based on such a complex 'reasoning' and 'deliberation'.***

Norm conformity as *routine* behavior

It is absolutely true that:

>> norm conformity and obedience 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:

Condition ==> Action;

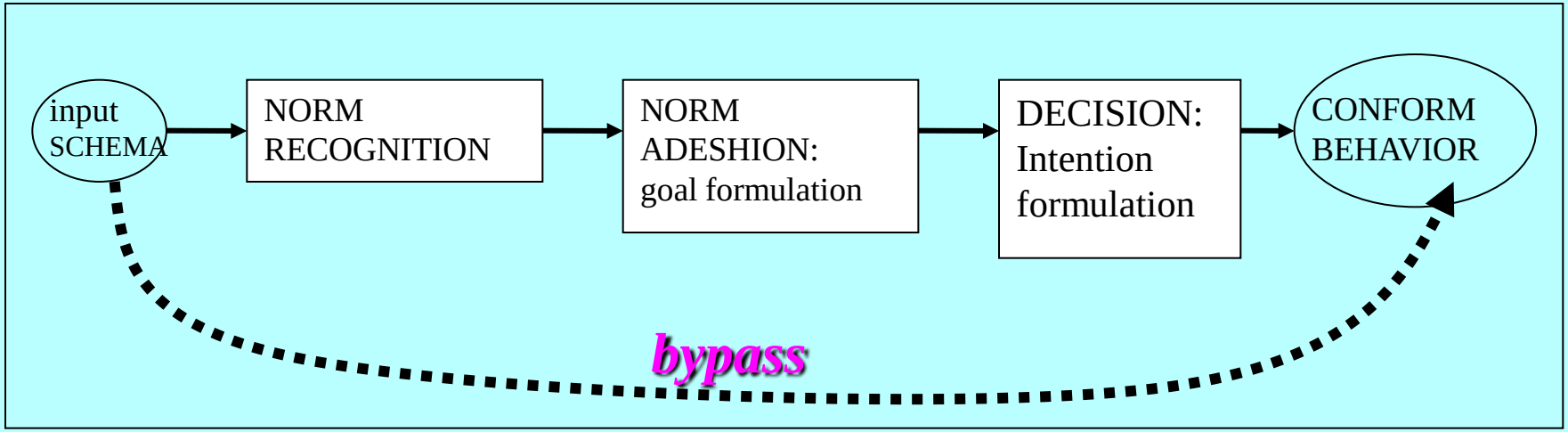
Recognized stimuli ==> Appropriate behavior.

Norm conformity as *routine* 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 be its origin and source, and *its cognitive background and justification.*

Norm conformity as *routine* behavior



7

The Issue

Goal -Adoption is how an autonomous agent is not an isle but becomes social, or better pro-social; that its (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?

From the Cognitive point of view

it is fundamental do not mix up:

(i) FUNCTIONS:
returns, outcomes that *maintain and reproduce* *a given behavior* (through reinforcement or selection)

and

(ii) mental GOALS:
returns, outcomes that are <u>foreseen</u> and <u>calculated</u>: we act *in view of* them and *instrumentally to* them

(c) The value of the adopted goal:

evaluation and "I care"

A necessary cognitive or learning condition for advanced form of goal-adoption (but also for its opposite: "aggression") is that

X is able to estimate <u>how much</u> the Goal of Y is valuable/important <u>for</u> Y.

In many cases, the value of the adopted goal for X is actually function of two variables:

(c) The value of the adopted goal:

evaluation and "I care"

the value of the adopted goal for X is actually function of two variables:

- ***How much is important for me the welfare of Y***, *her satisfaction, the fact that she realizes her goal;* let's call this *"how much a care of Y"* (not necessarily in altruistic or benevolent sense; also in commerce; for example: have I possible alternative partners if Y is discontent of me? if she will be unsatisfied?).

- ***How much I believe that G is important for Y; the value for Y****: is it a minor, marginal, goal for Y, or is very fundamental for her?*

Notice that this holds **also for an effective *aggression***, that is, for producing harms (for ex. in ***punishment***).

A harm actually is the frustration of a goal, the destroying of some good (already in possession or expected) of Y, of any kind.

I cannot really harm Y if I do not believe or learn that that is a value/goal for her.

How can I know that?

There are at least three ways for 'knowing' the value of a goal for the other (and that it is a goal of her):

> **Simulation (empathy, identification) and projection** of the simulated value in my mind.

Before this:
> From your behavior, used as a "cue" of the mental stuff. It is a rather simple heuristics: How much you look disposed to spend for obtaining (defending) G? The more you fight, invest, and are willing to work, the more G is valuable for you. (Of course, "to spend" is in broad sense: for example, how much you fatigue and fight with me for that bon? How much energy and time are you spending for it; how much you are disposed to risk for it?). For example, the longer and strongly you - an infant – cries, the stronger should be your hunger.

> From expressive behavior, which is signaling the intensity of your motivation: for example, rage and disposition to fight.

> A posteriori, from the expressive or behavioral consequences of your achievement or failure: sufferance, joy, reciprocation (and its entity), etc.

Without such an ***ability to appreciate the value of G <u>for you</u> (not for me)***
and – in advanced forms – doing this by a real mind reading, I would never be able to promise or threaten something to you.

Especially, **conditional promises or threats** aimed at inducing you to do or nor to do something.
I necessarily have to assume that what I promise or threat to you ***has for you more value/importance than the other goal of yours that I want suppress***.
Actually I try to influence you by creating or activating a conflict in you. *"If you finish your homework (if you stop to play) I will bring you to the movie"*. The manoeuvre is effective (and is rational) only if I believe and it is true that the **value** for you of *"going to the movie"* is greater than the value of *"continuing to play"* and of the effort of doing homework.

(d) Mind reading for coordination (Collaboration)

X has to understand whether Y is pursuing herself the G or not, and also whether G is part of a larger plan involving other actions of Y.

I have to understand from your behavior, structure or conditions that you are not able to achieve your goal. Or, if you are able and in condition to do the needed actions, I have to understand that you do not intend to do so (for some reason).

It is dangerous to *substitute* you: perhaps we will interfere with each other or duplicate the efforts. X should be sure at least that Y will realize that X intends to do or is doing the relevant action.

Analogously, if you have to do other actions in parallel or in sequence with my adoptive action, we must coordinate our actions. In those – frequent cases - it is a fundamental condition that Y knows that I have adopted her goal, or at least that I will do the action, or – minimally - that I'm doing or I have done it. In other terms, in those cases it is a necessary condition for an efficient adoption that Y 'delegates' that G/action to X, that 'relies' on X

X has to wonder about Y's possibility to observe his behavior or outcomes; X's action or its product actually are a 'cue' or a 'signal' for Y.

(b) Adoptive Goal vs. Adopted Goal

That I come to have your goal (and to pursue it), is not enough.
A simple mechanism of goal-transfer, of X coming to have Y's goal is not enough.

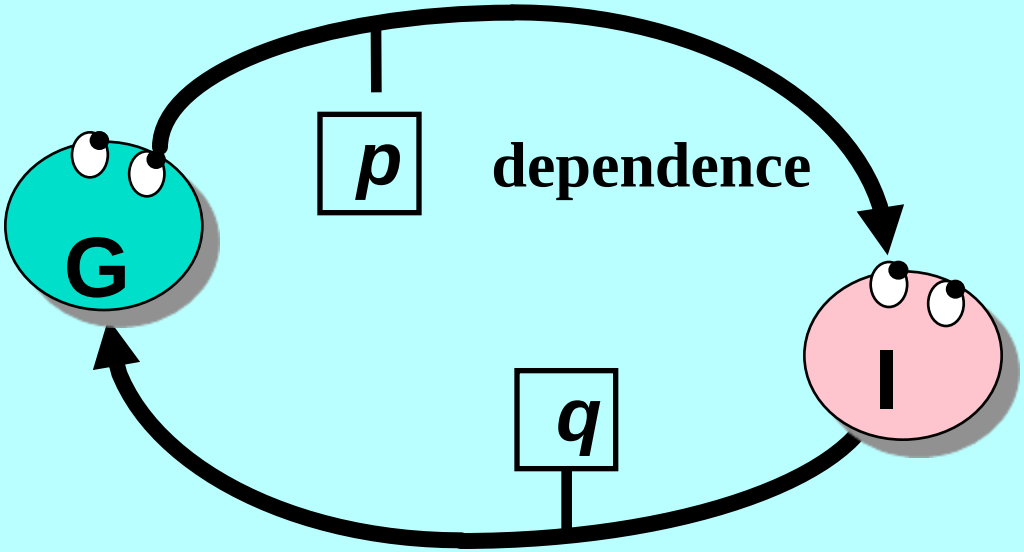
What is needed is the mental formulation of the "Adoptive Goal":
$Goal_x$ of Adopting $Goal_y$p (the "Adopted Goal").

X comes to have/pursue Goal that P (as his own goal: *(G_x p)*)
in order **Y realizes his goal**.

Given my goal that you realize your goal (that your goal – as your goal – be realized), *then* I get the goal that p.
This is real human goal adoption; not whatever mechanism such that if I believe that you wants P, then I want P.

Notice that the "Adoptive Goal" (the goal that you realize your goal) **is not "benevolence",** or **"altruism"** or **moral values etc. ('Social preferences')**
In fact not necessarily your achievement (welfare, satisfaction,...) is a terminal, non instrumental or selfish goal for my own calculated advantage.

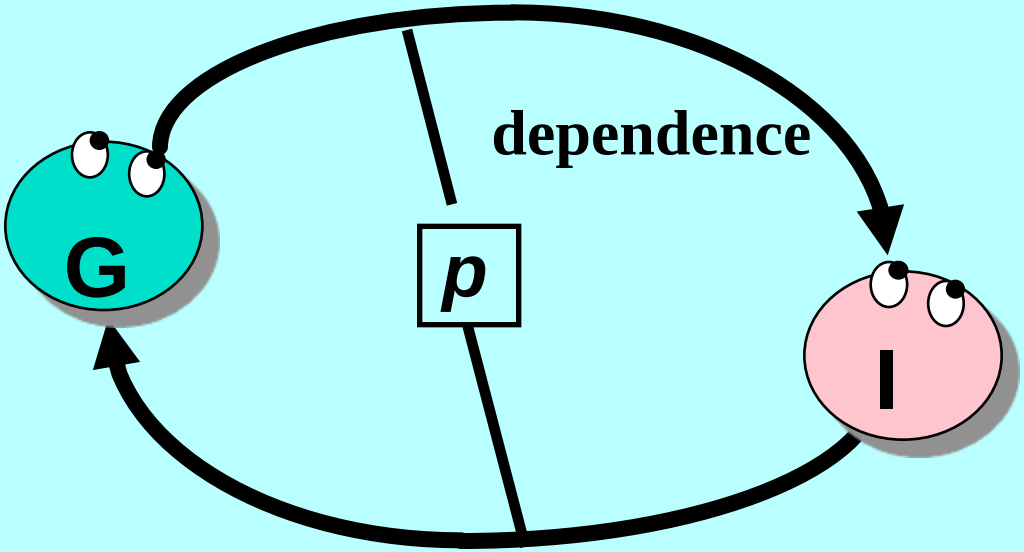
Reciprocal S-Dependence



exchange

cheating, defeating,
problems of reciprocation, ...

Mutual S-Dependence



strict **cooperation**

common goal, co-interested agents,
to defeat is self-defeating....

1.

Mind Reading
is for

‘influencing’ and ‘manipulating’

Influencing & Manipulating

Other Agents in the same world means: 'interference'

>> I can adjust my own behavior to the other's behavior(for *exploiting* or *avoiding* interference)

>> .... but, I can also CHANGE your behavior ('Influence'),

in order to obtain what I need or prevent harm.

In order to change your behavior

the best is changing your Mind:

- **I should 'understand' the hidden mechanisms regulating it;**
- **I should be able to act upon those 'mental' mechanisms: your goals, beliefs, preferences, intention, ...**

The special relevance of Goals and Motives for Social Autonomy

Autonomy as Self-Motivation

autonomy par excellence

because of the special role of **goals** in the definition of an interesting notion of 'agent'
A goal-autonomous Agent is an Agent endowed with its own goals.

an Agent is **fully socially autonomous** if:

- *It has its own Goals: endogenous, not derived from other Agents' will.*
- *It adopts goals from outside, from other Agents; it is liable to influencing. To be motivation-autonomous does not means to be autarchic or a-social; the agent can accept goals from others.*
- *It adopts other Agents Goals only if it sees the adoption as a way of enabling itself to achieve some of its own goals (i.e. the Autonomous Agent is a Self-Interested/motivated Agent).*

"self-interested" or "self-motivated" NOT EQUAL to "selfish"

"it own goals" NOT EQUAL to "egoistic goals"

Goal-ADOPTION

"Y has the Goal G1 since and until it is the Goal of X"

is NOT 'IMITATION'

is not "doing the same", "doing like the other"

It is doing something 'for' the other,
for realizing her Goal

Cooperation among humans is not just "behavioral"

Before being behavioral must usually be "mental":
***due to psychological representations, attitudes, decisions that teleologically guide behaviors*.**

Which are the mental representations and processes that mediate human cooperation?

We will examine only some (crucial) aspects of that problem: goal-adoption; goal-adhesion; expectations and prescriptions about the other mind and behavior.

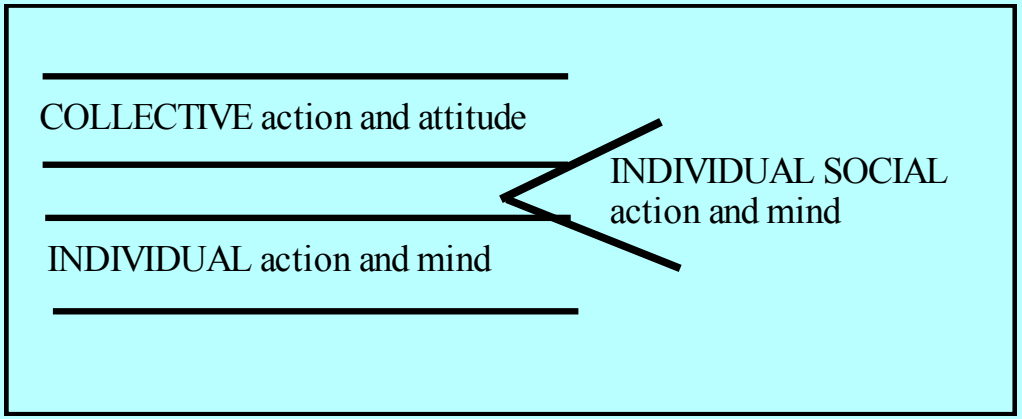
Social intelligence

"Social" in two senses:

a) **INTERACTIVE**:

b) **COLLECTIVE**

intelligence and problem-solving ability which either spontaneously or orchestratedly emerges from the interaction of (more or less intelligent) entities not able to solve a problem individually and thanks to their reasoning.

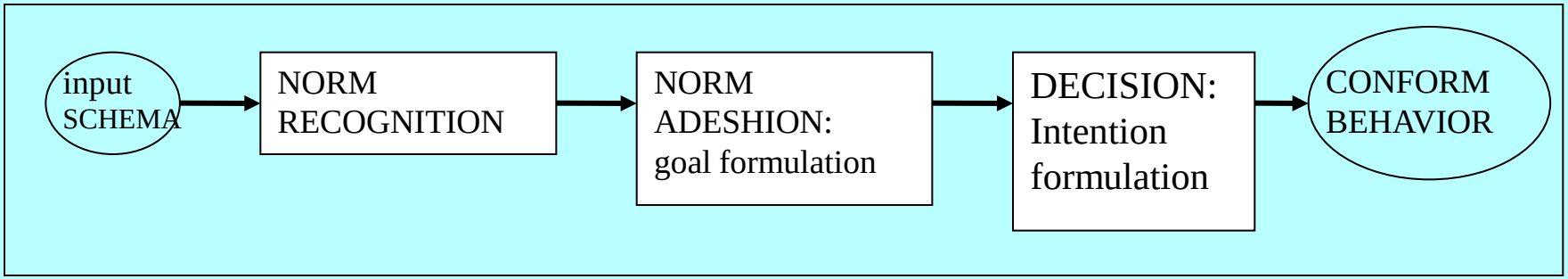


- ***SOCIAL ACTION WITHOUT A PLURAL SUBJECT***
- ***SOCIAL ACTION =/= COLLECTIVE =/= COMMUNICATION =/= PRO-SOCIAL***

From Goal-Adoption, decision, intention, … to ROUTINES

Norm conformity as *routine* behavior

Our quite rich cognitive characterization of the representations and processes underlying a behavior obedient to a norm, ……



……. shouldn't however give the idea of ***behavioral conformity as always based on such a complex 'reasoning' and 'deliberation'.***

Norm conformity as *routine* behavior

It is absolutely true that:

>> norm conformity and obedience 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:

Condition ==> Action;

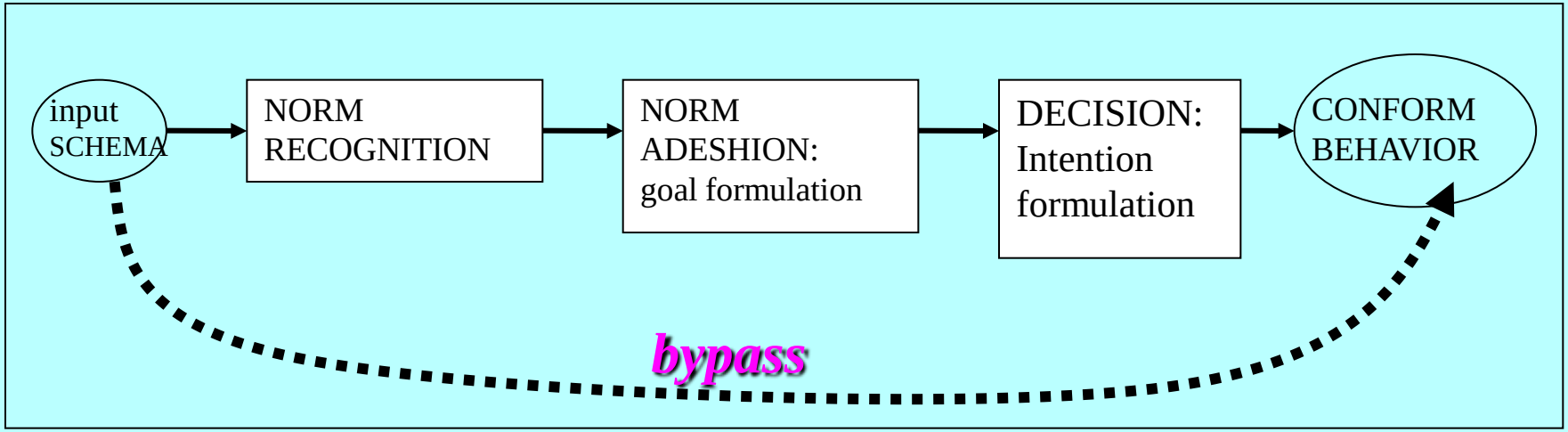
Recognized stimuli ==> Appropriate behavior.

Norm conformity as *routine* 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 be its origin and source, and *its cognitive background and justification.*

Norm conformity as *routine* behavior



7

A FEW WORDS ON "EXPECTATIONS"

Cognitive Anatomy *of* *Expectations* :

Expectations are ***compositional states*** *(and in part activities). Their ingredients are:*

- **a belief that *p*** about the future **(*prediction*) with its degree of certainty (%b) (how much one is sure that…);**

- **an epistemic goal (and activity): the goal to know, to check whether *p* happens to be true (“*expecting*”, “*waiting for*”)**

- **a *wish, desire, need,..* (in our vocabulary a generic “goal”, not in the sense of an objective to be pursued), with its degree of value (%v), of importance**

'Expectations' are not just 'Predictions'

We do not want to use 'expectations' (like in the literature) just to mean 'predictions', that is, epistemic representations about the future.

We consider, in particular, a 'forecast' or 'prediction' as a mere belief about a future state of the world

For us 'expectations' have a more restricted meaning

(and this is why computers can produce weather 'predictions' or 'forecasts' but do not have 'expectations')

Cognitive conditions for Goal-Adoption

>> *Beliefs about Y's goals (Mind reading)*

A fundamental condition is an intentional stance of X (the adopter) towards Y (the adopted guy), and more precisely – if Y is considered a cognitive agent – ***a mind-reading attitude in X towards Y.*** X has to ascribe to Y a given internal goal (of any kind)

Bel x (Goal y p)

and X decides to "appropriate" that goal, since and until it is the goal of Y.

So X **comes** to have the same goal:

(Goal x p)

but ***relativized*** to that belief.

>> *The value of the adopted goal: arriving to an adoptive INTENTION*

Goal Adoption is not enough. **I can adopt a goal of yours, take it into account, but then decide to ignore it**, to violate your request, right or expectation (if it is the case), because I have more important or urgent goals of mine.

So, to have an *adoptive behavior*, **it is necessary to arrive to a corresponding 'adoptive intention', that is, the intention corresponding to the adoptive goal.**

If I adopt your goal that p, I have to eventually formulate the *intention "that p"* and thus the *intention "to do"* something to bring about p.

Kinds & Motives of Goal-Adoption

Goal-Adoption

is not 'benevolence' or 'altruism'

..... *Social Preferences*

there are various Motives
for doing something *for* the others

there are various Kinds
of Goal-adoption

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."* (A. Smith, *An Inquiry into the Nature and Causes of the Wealth of Nations*, 1776)

In Smith's perfect description of exchange <u>in merely selfish terms</u> it is clear that there is non benevolence or altruism 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: for 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.

NORMS as *MULTI-AGENT* artifacts

In all those "minds" the N is an imperative on a class of agents.

However, in I and W' 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.

Let's go deeply in S's mind and adoptive process

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 norms;**
- **to issue the norm as a norm (that is "counting as" a norm; recognizable as a norm, 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).**

Internalization (and why it matters)

Moral 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!", "Don't 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"

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;…

Norms exploit and count on

a special process/kind of Goal-Adoption

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

The N wants to be followed for an internal mechanism reflecting it; for a Goal of following the N.

>> **They prescribe also a "mental attitude".**

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, 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".**

Our general PERSPECTIVE

The *"Cognitive Mediators"* of Social Phenomena

Social phenomena are due to the agents' behaviors, but…
the agents' behaviors are due the the *mental mechanisms* controlling and (re)producing them. (Castelfranchi, Conte, Miceli, Falcone,…)

• How the norm should ***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 norm from a mental - not just a behavioral - point of view? What does it mean to "obey"?

• What kind of mental attitude the Norm "prescribes" to, builds into, the agents?

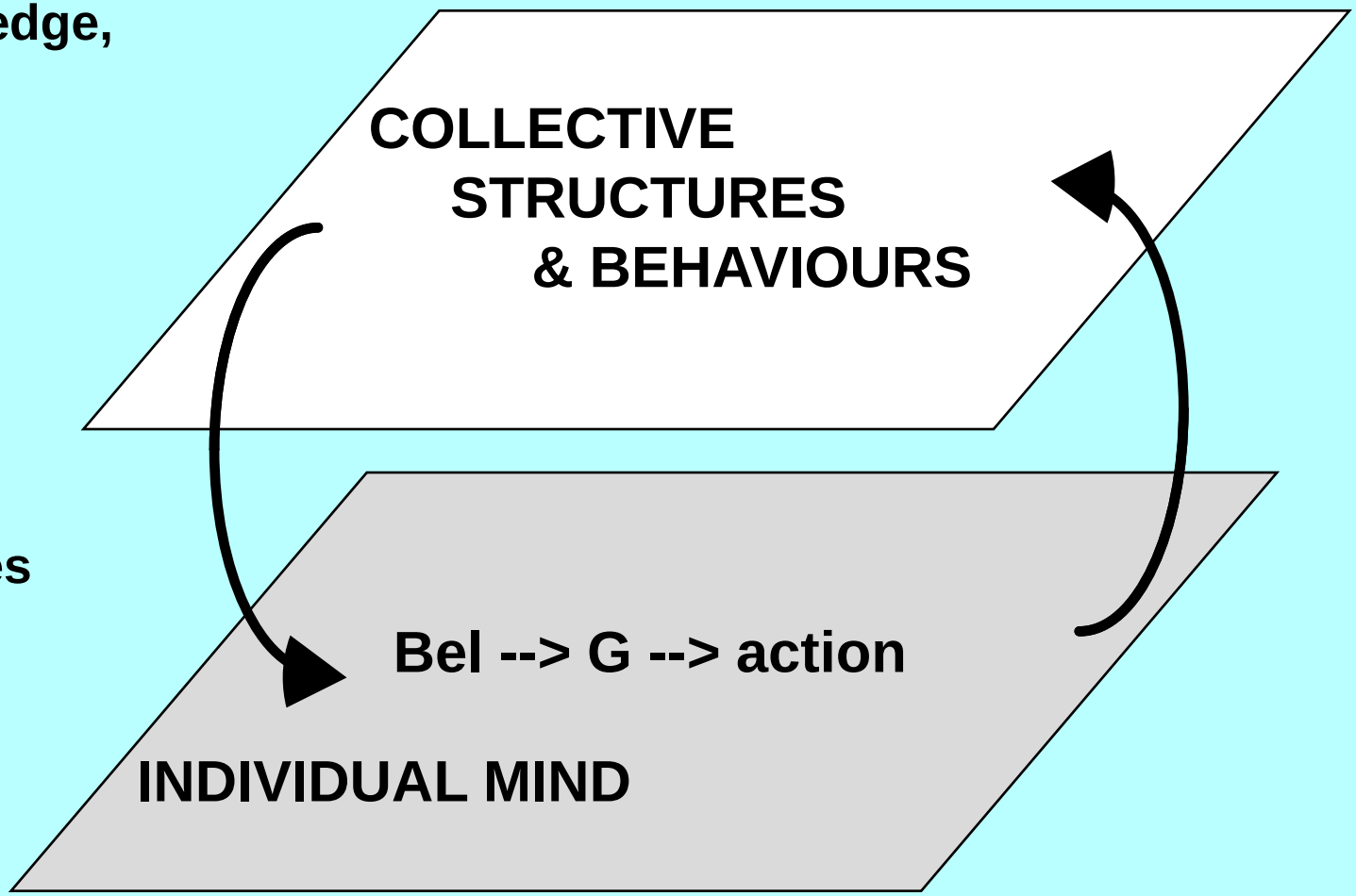
Mind is not enough

emergence

**not only knowledge,
mutual beliefs,
reasoning,
shared goals**

and

**deliberately
constructed
social structures
and
cooperation**



Cognitivizing "Norms"

Our theses about Norms (Conte & Castelfranchi) **were that:**

> 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"*); differently from other social phenomena like *social functions*.

> they require different and complementary "roles" with their specific "minds" or "mental attitudes": the *subject* of the N, the *watcher*, the *issuer*.

> we explained why a *subject* also becomes a *watcher* and an (implicit) *issuer*.

> they are based on a specific process of Goal-Adoption or better Adhesion; since they have the nature of an "imperative"

> they are aimed at being "obeyed" for specific motives: not for external rewards, not for benevolence, etc. but for the recognition of values, authority, ...

From *BELIEFS* (about the others' behavior and mind)

to *GOALS* (about the others' behavior and mind)

An ideal path:

By **"Prescription"** I mean not just a *wish* about our behavior, or a goal (to do something).

Not only your goal about our behavior as a "request" (implicitly or explicitly) *communicated* to us, in order to elicit "adhesion",

>> but, **it is an *entitled* request, that *wants to be recognized as entitled* and accepted *because is entitled*.**

True Norms (social, moral, formal and legal) ***have a "prescriptive" nature. They are aimed at be "obeyed": want our behavior for specific reasons, with a specific mental attitude of us.***

Generalized Goal-Adoption

- 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 punish Z.**

.