Available online at www.sciencedirect.com
SCIENCE DIRECT
Cognitive Systems Research 7 (2006) 209–219
Cognitive Systems RESEARCH
www.elsevier.com/locate/cogsys

Argyll-Feet giants: A cognitive analysis of collective autonomy

Action editors: Luca Tummolini and Cristiano Castelfranchi

Rosaria Conte *, Paolo Turrini

Laboratory of Agent-Based Social Simulation, Institute of Cognitive Science and Technology, National Research Council, Via San Martino della Battaglia 44, 00185 Roma, Italy

Received 10 April 2005; accepted 7 November 2005
Available online 30 March 2006

Abstract

In this paper, a formal theoretical analysis of the notion of collective autonomy is proposed. Collective agents are defined as *multi-agent strategies* for goals not necessarily shared by the agents that are interdependent in achieving them. Two previously defined notions of autonomy (plan- and goal-autonomy) are applied to the collective level. Internal collective autonomy is distinguished from external, and the effect of collective internal autonomy is analyzed in terms of influence exercised on the collective's members.

Keywords: Theory of organizations; Multiagent systems; Autonomy

1. Introduction

In this work, a notion of collective autonomy will be presented and discussed.

One could question whether investigating collective autonomy makes any sense at all, as no consolidated notion of collectives is yet available. However, autonomy per se is a heuristic ingredient for theorizing upon collectives, and because of its relevance in multi-agent decision-making it is an issue of concern within the study of collective intentionality and cognition.

On the one hand, in fact, the cognitive perspective has a heuristic potential for the investigation of collective autonomy, lending it its own conceptual, theoretical and formal instruments of analysis. As will be shown throughout the paper, the investigation of the issue at hand largely benefits from the extension to the collective level of mental states such as goals, beliefs, etc. and from the formal models used for describing and treating them.

On the other hand, a theory of collective autonomy bears important consequences for the study and application of intelligent systems. This is so not only because autonomy is a property essential, though not exclusive, of intelligent systems, but also because the interrelationships between individual and collective autonomy are based upon, and mediated through, mental states.

In our opinion, collective autonomy needs to be specified at various levels, within and without the system, and the interrelationships between the system's autonomy and that of its members need to be explored.

Hence, we will turn to the theory of Agents and Multi Agent Systems (AT and MAS), where the notions of autonomy and collective agents have been separately provided. Each notion is relevant for highlighting important phenomena (different types of systems, coordination and cooperation, organisations, etc.) and their intersection is essential for investigating the question as to when and how supra-individual systems become "ends in themselves", although, to our knowledge, the issue of autonomy at the level of collective agents has not yet been explicitly addressed in the field of MAS or AT.

The present work is based on the following claims:

- The property of autonomy is crucial for characterising not only different types of individuals, but also different types of multi-individual agents. Ontological models

* Corresponding author.
E-mail address: rosaria.conte@istc.cnr.it (R. Conte).

1389-0417/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.cogsys.2005.11.011

(for instance, Guarino & Welty, 2000, 2002) have shown that understanding emergent properties helps classifying subtypes at the individual level.
- Social systems, institutions, organisations, etc. vary in their degree of autonomy, both at the supra-individual and at the individual level: autonomy is not a none-or-all notion but a gradual one.
- The level and degree of autonomy of supra-individual systems is useful to both understand and optimise organizations. Indeed, collective autonomy lends itself to highlight a number of crucial theoretical issues in social and political science, for example, decentralised vs. centralised forms of organisation can be compared. Also, a model of collective autonomy is essential for the enforcement of coalition, groups, teams, etc. In particular, it is important to establish the advantages and disadvantages of more or less autonomous groups of agents, hence to design artificial multi-agent systems endowed with the desirable level of autonomy before assigning them some task.

The paper is organised as follows. The notion of autonomy on the one hand, and that of collective on the other, will be discussed separately, referring to definitions introduced in previous works (Castelfranchi & Falcone, 2004; Conte & Castelfranchi, 1995). Secondly, these old concepts will be reunited in a definition of collective autonomy. Thirdly the potential of the present analysis in highlighting questions for future research will be shown.

Finally, in order to formally handle notions like autonomy and supraindividual properties we need to translate these within logical structures that allow dealing with features of interaction and emergence.

Thus, ATEL logic (Van der Hoek & Wooldridge, 2003) will be empowered to describe the fundamental ingredients of the model, whereas, to account for relations between different kinds of autonomy, a formal system for ontologies expressed with first-order logic will be adopted.

2. The logic model of reference

A brief description of the logic used will be provided. This will mainly concern the core system, which will be an extension of ATEL, an epistemic propositional temporal logic elaborated by Wooldridge and Van der Hoek. Higher order notions – like for instance the growth of autonomy in relation with other factors – will be analyzed simply switching to a first-order variant. What follows is a brief presentation of ATEL, and a full description of the syntax, semantics, axioms and inference rules used throughout the paper.

ATEL is a alternating-time epistemic version of CTL logic that combines path quantifiers (“necessary” and “possible”), with tense modalities (“always” and “eventually”). Moreover, ATEL is agent-based: it is defined for primitive objects that are called agents, which are in charge of the transitions between world states. Path quantifiers in ATEL are expressed as cooperation modalities: $\langle\langle \Gamma \rangle\rangle \Diamond \varphi$ means that agents in Γ can cooperate to eventually ensure that φ holds, whereas $[\![\Gamma]\!] \circ \neg\varphi$ means that agents in Γ cannot cooperate in order to (cannot avoid) to see to it that at the next state $\neg\varphi$ holds. $[\![x]\!]$ and $\neg\langle\langle x \rangle\rangle\neg$ are duals. The original version of ATEL is knowledge-based. We choose not to deal with knowledge but with beliefs (therefore we are not going to use S5 axiomatic) and to introduce goals in the system.

Syntax. The set of well-formed formulas is defined by:

$$\varphi, \psi, \ldots ::= p, q \ldots . | T | \neg\varphi | \varphi \wedge \psi | \langle\langle \Gamma \rangle\rangle \Diamond\ \varphi | \langle\langle \Gamma \rangle\rangle \circ \varphi | \langle\langle \Gamma \rangle\rangle \Box\ \varphi | \langle\langle \Gamma \rangle\rangle \varphi U \psi | \mathrm{Bel}_a \varphi | \mathrm{Goal}_a \varphi$$

Semantics. We define an Alternating Epistemic Transition system as a tuple $\langle \Pi, \sum, Q, \mathrm{Bel}_1, \ldots, \mathrm{Bel}_n, \mathrm{Goal}_1, \ldots, \mathrm{Goal}_n, \pi, \delta \rangle$ where:

- Π is a finite, non-empty set of atomic propositions;
- $\sum = \{a_1, \ldots, a_n\}$ is a finite, non-empty set of agents;
- Q is a finite, non-empty set of states;
- $\mathrm{Bel}_a \leqslant^{1} Q \times Q$ is an epistemic accessibility relation for each agent $a \in \sum$;
- $\mathrm{Goal}_a \leqslant Q \times Q$ is a volitional accessibility relation for each agent $a \in \sum$;
- $\pi : Q \to 2^{\Pi}$ gives the set of primitive propositions satisfied in each state;
- $\delta : Q\ x \sum \to Pw(2^{Q})$ is the system transition function, which maps states and agents to the choices available to these agents. The system is completely determined by its component agents: taken $\delta(q, a)$ the set of choices available to a in q, for every $q \in Q$ and every $Q_1, \ldots, Q_n$ of choices $Q_a \in \delta(q, a)$, the intersection $Q_1 \cap \ldots \cap Q_n$ is a singleton.

Furthermore, two fundamental notions of computation and strategy have been introduced:

- *Computation*. A state q' is an a-successor of a state q if there exists a set $Q' \in \delta(q, a)$ such that $q' \in Q'$. succ(q, a) is the set of a-successors of q, which is simply a successor if this holds for the set of all agents ($\sum$). A computation on a tuple AETS is an infinite sequence of states $\lambda = q_0, q_1, \ldots$ such that for all $u > 0$, the state q_u is a successor of q_{u-1}. A q-computation starts from q. $\lambda[u]$ is the uth state in λ. $\lambda[0, u]$ and $\lambda[u, \infty]$ are the finite prefix $q_0, \ldots, q_u$ and the infinite suffix $q_u, q_{u+1}, \ldots$ of λ, respectively.
- *Strategies*. A strategy is an abstract model for decision making. It is a sort of plan. A strategy f_a for an agent $a \in \sum$ is a total function $f_a : Q + \to 2^{Q}$ which must satisfy the constraint that $f_a(\lambda \cdot q) \in \delta(q, a)$ for all $\lambda \in Q^*$ and $q \in Q$. Given $\Gamma \leqslant \sum$, $F_\Gamma = \{f_a | a \in \Gamma\}$, one for each agent in the group. We define out(q, F_Γ) to be the set of possible outcomes if every agent a follows the corre-

[1] To be read as “included in”.

sponding f_a strategy starting from $q \in Q$. The set of all agents can cooperate to uniquely determine the system: out(q, $F_{\sum}$) is a singleton.

Interpretation

- $S, q \models T$;
- $S, q \models p$ iff $p \in \pi(q)$ (where $p \in \Pi$);
- $S, q \models \neg\varphi$ iff $S, q\neg \models \varphi$;
- $S, q \models \neg\varphi \wedge \psi$ iff $S, q \models \varphi$ and $S, q \models \psi$;
- $S, q \models \langle\langle \Gamma \rangle\rangle \circ \varphi$ iff there exists a set of strategies F_Γ, one for each $a \in \Gamma$, such that for all $\lambda \in \text{out}(q, F_\Gamma)$, we have $S, \lambda\,[1] \models \varphi$;[2]
- $S, q \models \langle\langle \Gamma \rangle\rangle \Box\, \varphi$ iff there exists a set of strategies F_Γ, one for each $a \in \sum$, such that for all $\lambda \in \text{out}(q, F_\Gamma)$, we have $S, \lambda[u] \models \varphi$ for all $u \in$ Naturals;
- $S, q \models \langle\langle \Gamma \rangle\rangle\, \varphi U \psi$ iff there exists a set of F_Γ, one for each $a \in \sum$, such that for all $\lambda \in \text{out}(q, F_\Gamma)$, there exists some $u \in$ Naturals such that $S, \lambda[u] \models \psi$, and for all $0 \leqslant v < u$, we have $S, \lambda[v] \models \varphi$;
- $S, q \models \text{Bel}_a\varphi$ iff $\forall q' \in \text{Bel}_a(q)$ it holds that $S, q' \models \varphi$;
- $S, q \models \text{Goal}_a\varphi$ iff $\forall q' \in \text{Goal}_a(q)$ it holds that $S, q' \models \varphi$.

Axiomatic. We are going to use all propositional and predicate logic validities, and to give axiomatic definitions as we go. For inference rules, modus ponens and modal generalization will be adopted.

Functions and relations will be defined as $f\{n, m\}(x_1, \ldots, x_n)$, $A\{p, q\}\{(y_1, \ldots, y_q)$ meaning respectively the mth function f with n positions and the pth relation A with q positions: $x_1, \ldots, x_n$ and $y_1, \ldots, y_q$ are terms.

From now on we use indifferently the notation $s_0, \ldots, s_n$ or $q_0, \ldots, q_n$ for world states, M or S for models. When dealing with a few variables we may happen to distinguish them in various ways: for instance x', x'', $x'''\ldots$ or p, q, $r\ldots$ The 'kind' of variable they belong to will be for the sake of clarity always specified.

3. Limited autonomy

Needless to say, the property of autonomy has played a foundational role in Agent Theory. There are several notions of autonomy. Aside from general agent definitions, most people focus on an autonomous agent as: "...a system situated within and a part of an environment that acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future" (Franklin & Graesser, 1996). Far from a property of agents in isolation, autonomy characterises *situated* agents. Less convincingly, Hexmoor (2004) conceives of an autonomous agent as an agent beyond any other agent's control. Much alike other physical entities, such as the climate, autonomous agents would be uncontrollable entities.

In this paper, however, a different perspective on autonomy is adopted. Rather than uncontrollable or unpredictable entities, autonomous agents are defined as systems endowed with the capacity to choose whether to have some world-state as a goal. Obviously, however, such a capacity is inherently limited (see also Castelfranchi & Falcone, 2004), in at least two senses:

- Autonomy is a *relational* notion; hence you may be autonomous with regard to a given set of goals but not with regard to all of your goals, and before a given set of agents but not before all of the agents you may interact with. Therefore, the predicate AUTONOMOUS (x, Γ, P) will be defined for three terms, where x is the agent that is autonomous, Γ is the set of agents from whom x is autonomous, P is the set of goals (or plans) that restrict x's autonomy from Γ.
- We will show that P cannot coincide with the whole set of goals an agent can autonomously generate.
- Autonomous agents undergo *social influence*. An autonomous agent is one that may receive external inputs and even accept them, provided such an acceptance is the result of a decision grounded upon the agent's internal criteria. For instance, it may happen that I believe that (a) another agent wants to transmit a new belief to me, that (b) he is not lying and that (c) he is reliable about that belief, and still I may not adopt (for some reason) the other's belief: $M, s_0 \models \text{Bel}_y\text{Goal}_y\text{Bel}_x\varphi \wedge \text{Bel}_x\text{Bel}_y\varphi \wedge \text{Bel}_x\,\text{Reliable}(y, \varphi) \wedge \text{Goal}_x\text{Bel}_x\neg\varphi$ implies $M, s_1 \models \text{Bel}_x\neg\varphi$ given s_1 is x-goal-accessible from s_0.[3]

Hence, an autonomous agent can be influenced, coerced, even manipulated, but only via its decision-making. A person that, under gunshot, gives away her wallet is still an autonomous system, although to a fairly limited degree. The extorter succeeded in modifying the victim's goals because he managed to be credible enough: he had to predict the decisions of the victim, and provide her with good reasons for yielding.[4] This shows that coercion does not exclude autonomy, although it reduces it, and that autonomy has nothing to do with predictability. A chaotic system like the weather does not make decisions. On the other hand, systems can take fairly predictable decisions, as in the case seen above, and still be autonomous. However, it is true that systems are autonomous when they "could have behaved otherwise": a complicated way to say that they have had a choice.[5]

[2] Which is equivalent to $\neg[\![\Gamma]\!] \circ \neg\varphi$, to be read as "it is not true that the agents in Gamma cannot avoid that at the next state not-phi will hold".

[3] If we define Bel_x Reliable(y, φ) as $\text{Bel}_x(\text{Bel}_y\varphi \rightarrow \varphi)$, by applying modus ponens with $\text{Bel}_x\text{Bel}_y\varphi$ we get $\text{Bel}_y\varphi$ and we may get a contradiction. In fact to solve these discrepancies we can think of beliefs as mental objects having a given strength.

[4] In fact the coercion does not modify the rule by which an agent adopts a goal but it strengthens the belief of the incoming danger.

[5] This puts forward an interesting link between autonomy and responsibility, saying that only autonomous agents can be in fact responsible.

4. Autonomy and intelligence

If autonomy is limited, intelligent systems are not always autonomous. Humans, for example, are only partially autonomous. If I get drugs or other substances injected (Conte & Castelfranchi, 1995), my perceptions and my goals will be altered and distorted against or independent of my will. In principle, systems can undergo two different types of influence:

- *Mind-off*. Intervention that either reduces the entity's capacity to act (immobilizing) or alter its physiological and neurological processes (drug injection, surgery, etc.) and as a consequence its mental properties. Only the former type of influence is compatible with autonomy. Hence, systems undergoing both types of influence are only limitedly autonomous. In ATEL, we would say that we allow agents to have strategies that can directly modify others' mental states: $M, s_0 \models \langle\langle x \rangle\rangle \circ \mathrm{Bel}_y \varphi$ (the agent x has a strategy to see to it that at the next state y believes that phi, independent of y).
- *Mind-on*. Influence is exercised through the entity's decision-making, in ATEL we would say that we do not allow agents to have strategies that can directly modify others' mental states, that can happen only by means of an autonomous response: ($M, s_0 \models \mathrm{Goal}_x \mathrm{Bel}_y \varphi$ and for all λ, M, $\lambda[1] \models \mathrm{Bel}_y \varphi$) iff $M, s_0 \models \langle\langle y \rangle\rangle \circ \mathrm{Bel}_y \varphi$ (y is autonomous iff for any world state φ, x wants y to believe y will believe it only if it y has (and execute) a strategy to see to it that y believes that φ).[6]

  The modification of its current goals is due to a *rule of goal-generation*:
  If a given entity x is autonomous with regard to a world-state q', q' will become a goal of x if there is at least one world-state q'' that x wants to be realised such that x believes q' to be instrumental for q''.

To formalize this rule we need to define the notion of instrumentality. We say that a state of the world is a means for another if and only if there is a strategy that allows an agent to eventually bring about the second, being true the first. $\models \text{MEANS-FOR}(q', q'')$ iff $\exists x \in \sum$ such that $M, s' \models q' \rightarrow \langle\langle x \rangle\rangle \Diamond q''$.[7]

Then we say that if q'' is wanted and q' is believed to be means for q'' then q' can be adopted as goal.[8] $M, s' \models \mathrm{Bel}_x$ MEANS-FOR $(q', q'') \wedge \mathrm{Goal}_x q''$ implies that $M, s' \models \langle\langle x \rangle\rangle \circ \mathrm{Goal}_x q'$.[9]

We can state the correlation between autonomy and goal generation by saying that if an agent is autonomous before a set of agents regarding a state of the world, then it can apply a goal generation rule. And if he can apply a goal generation rule concerning a state of the world then he is autonomous before that set of agents for that sw.

It follows that: $M, s' \models$ AUTONOMOUS(x, Ω, q'') iff for all λ, M, $\lambda[1] \models \mathrm{Goal}_x q''$ and $\exists q' \neq q'' \in \Pi | M, s' \models \mathrm{Goal}_x q' \wedge \mathrm{Bel}_x$MEANS-FOR $(q'', q') \wedge \mathrm{Goal}_x \mathrm{Goal}_x q'' \wedge \langle\langle x \rangle\rangle \circ \mathrm{Goal}_x q'' \wedge [\![\Omega]\!] \circ \mathrm{Goal}_x q''$, where $\Omega \leqslant \sum$, it is a subset the set of agents. In words, an agent x is autonomous with respect to a set Ω for a goal q'' if and only if it is going to have q'' as goal and it exists a q' different from q'' that x has already as a goal of his', believes q'' to be a means for achieving q', has the goal to have the goal that q'' and a strategy to generate it, and no agent in Ω can do anything to avoid it.

Notice that a full autonomy is empirically unattainable in nature, and that criteria for generating goals include self-interest, biological fitness, moral standard, etc.

Therefore, it is possible to show that agents have to start from a non-empty set of hardwired goals, that is a goal they do not adopt by means of GGR. Such a set in fact, enables GGR: it is not possible to apply such rule from an empty set of goals.

Theorem 1. *We cannot have an autonomous agent with only one goal.*

Proof. Let us proceed by induction on the number of goals of an agent and let GOALS(x) denote this set. Suppose #GOALS(x) = 1, then suppose the only goal is $\{p\}$ and the agent is autonomous about it. Then for GGR, $\exists q \neq p \in \Pi \|$ for all λ, M, $\lambda[1] \models \mathrm{Goal}_x p$ and $M, s_0 \models \mathrm{Bel}_x$ MEANS-FOR$(p, q) \wedge$ $\mathrm{Goal}_x \mathrm{Goal}_x p \wedge \langle\langle x \rangle\rangle \circ \mathrm{Goal}_x p \wedge$ $[\![\Omega]\!] \circ \mathrm{Goal}_x p$. But this is false in our case, and so is the biconditional. Thus, we cannot have an autonomous agent with only one goal.

It has to be noted that the all set of goals must be non-empty. And since, the very notion of agency implies the notion of goal-driven systems. To have agents, we actually need to have goals.

Our theorem claims that an agent starts to be autonomous only by acquiring goals, starting from hardwired

[6] Many forms of belief induction happen with no cooperation at all. For instance, sometimes we simply cannot avoid to perceive a set of points as a triangle instead of something else (i.e., gestaltic shapes). Anyway it is possible to think of perception as result of an active process agents carry out. Cooperation does not happen in manipulation, as well, where agents make believe something without the other believing to be manipulated. We can model this kind of interaction by adding the fact that an agent has a strategy not to let the other believe that he had not a manipulative goal: $M, s_0 \models \langle\langle x \rangle\rangle \circ (\langle\langle x, y \rangle\rangle \circ \mathrm{Bel}_y \varphi \wedge \mathrm{Bel}_y \varphi \neg \mathrm{Goal}_y \mathrm{Bel}_y \varphi) \wedge \mathrm{Goal}_y \mathrm{Bel}_y \varphi$.

[7] We can have necessary means for reaching goals: something without which a goal cannot be reached. This is expressible with a biconditional. $M \models$ N-MEANS-FOR(q', q'') iff $\exists x \in \mathrm{Agt}$ and $\exists s' \in S$ such that $M, s' \models q' \leftrightarrow \langle\langle x \rangle\rangle \Diamond q''$. It follows that $\models$ N-MEANS-FOR$(q', q'') \rightarrow$ MEANS-FOR (q', q'').

[8] This is a deductive kind of goal generation rule: agents adopt goals ascribing states to the class of means. In fact for a state to be ascribed to a class we need also induction and reasoning on properties.

[9] Such a language does not properly account for causation, which is not material implication. Nevertheless it introduces the notion of "getting to have a strategy" enabled by a particular state of the world. But lots of phenomena seem not to be captured. "If it rains I have a strategy to eat many mushrooms". If I believe rain cannot be inducted by my behaviour, does it make sense to say that I have the goal "that it rains"?

ones. In fact only internal processes can guarantee autonomy. □

Corollary. *Any agent has at least a hardwired goal. Thus, no agent is completely autonomous.*

Proof. Suppose Δ is the set of goals an arbitrary agent a has. If Δ amounts to one goal, that goal is hardwired (for the Theorem proved above).[10] If it amounts to more, then there exists at least a goal that constitutes the basis for goal generation (for the same theorem), that is a hardwired one. If $\#\mathrm{GOALS}(x) = 2$ either they are both hardwired or one only has been autonomously generated: as shown before, it cannot be that both are. This properties holds for $\#\mathrm{GOALS}(x) = n + 1$. □

Internal processes include mental and physical maturation as well as reasoning and deliberation. Thanks to maturational processes, autonomous agents spontaneously drop some old goals or acquire new ones or simply find some of them modified. Thanks to reasoning, autonomous agents decide whether to include a certain world-state among its goals or not. Such a world-state may be provided from the outside (it may be a request, a command, etc.), but it may also be suggested from agent's own experience. Maturation is a fundamental source of goal updating. Indeed, a model of autonomous agents ought to account for such a process. However, in the present paper we will limit ourselves to account for deliberate goal-generation. As we will see later on in the paper, this aspect plays a more crucial role in the collective sphere than the former process does. However, autonomy is limited in many senses. First of all, systems, including humans, may undergo external influence and this may be directed to modify their behaviours either bypassing their decisions (as is the case with, say, strictly physical coercion[11] or drug injection) or acting on their decision-making (by means of cognitive influencing, intimidation, incentive or persuasion). Finally, properties arising in social interaction, such as responsibility and accountability, affect agents' decisions both directly and indirectly, modifying others' requests and expectations.

In Conte and Castelfranchi (1995), two levels of autonomy have been distinguished: plan-autonomy and goal-autonomy, which are of some importance also in the case of collectives.

- *Plan-autonomy*. The system's capacity to decide how to achieve a hardwired goal, which does not include the capacity to select external inputs and modify one's own goals in terms of own criteria.
  - In this case $M, s' \models$P-AUTONOMOUS (x, Γ, q'') iff $M, s' \models \mathrm{Goal}_x q'' \wedge \langle\langle x \rangle\rangle \circ \langle\langle x \rangle\rangle \Diamond q''$, where x does not belong to Γ. In words, plan autonomy is the capacity to generate a strategy to achieve one given goal.
  - In case of strong plan autonomy (that excludes goal autonomy) then $M, s' \models \mathrm{sP\text{-}AUTONOMOUS}(x, \Gamma, q'')$ iff $M, s' \models \mathrm{Goal}_x\ q'' \wedge \langle\langle x \rangle\rangle \circ \langle\langle x \rangle\rangle \Diamond q'' \wedge [\![x]\!] \Box \mathrm{Goal}_x q''$, in where $[\![x]\!] \Box \mathrm{Goal}_x\ q''$ means as in ATEL that the agent x cannot avoid that it always has a goal q''. Informally, strong plan autonomy add the fact that there is no alternative to have a given q'' as goal.

The latter case might be of interest in intelligent software agent applications, which cannot help complying with users' requests, but might be requested to be able to plan autonomously for executing them, as well as to reject the requests of others.

- *Goal-autonomy*. Modify, acquire and drop own goals by means of decisions based on internal criteria.
  - In case of Goal dropping. $M, s' \models G$-AUTONOMOUS $(x, \sum, q')$ if $M, s' \models \mathrm{Goal}_x q' \wedge \langle\langle x \rangle\rangle \circ \neg\, \mathrm{Goal}_x q'$.
  - In case of Goal acquisition. $M, s' \models G$-AUTONOMOUS $(x, \sum, q')$ if $M, s' \models \langle\langle x \rangle\rangle \circ \mathrm{Goal}_x q' \wedge \neg\, \mathrm{Goal}_x q'$ Informally, goal autonomy towards the set of all agents is the capacity to drop or to acquire a goal.
  - As these are the two fundamental cases, combining the two previous notions, we have a full definition of autonomy. $M, s' \models$ AUTONOMOUS $(x, \sum, q')$ iff $M, s' \models \langle\langle x \rangle\rangle \circ \mathrm{Goal}_x q' \wedge \neg\, \mathrm{Goal}_x q'$or $M, s' \models \mathrm{Goal}_x q' \wedge \langle\langle x \rangle\rangle \circ \neg\, \mathrm{Goal}_x q'$.

5. Supra-individual action

In substance, an action is supra-individual in the fullest sense when it must be accomplished by two or more agents.[12] It is a rigid property (in the sense of Guarino-Welty, 2002), as we cannot have a supra-individual action carried out by one agent.

There are several subtypes of such actions (each such subtype is respectively constituted by tokens), depending on two dimensions, namely the agents sharing the same goal, and their interdependence in the pursuit of it:

[10] We do not consider goal dropping, that is the fact that a goal can be discarded as time goes by. Such theorem anyway is static in time, and considers goals agent have and have had. Generalization to time can be easily done, weakening the theorem while considering it. We also absume generation is neither reflexive (goals are not generated from themselves), nor symmetric (it cannot be given the case that q is generated from p and p from q), but transitive (if p is generated from q and q from s, then p is indirectly generated from s). In this theorems we are implicitly assuming that GGR is the only rule we have to autonomously generate new goals.

[11] It is important to notice that physical coercion may also be a means for influencing the victim's decisions: a robber asking for you wallet while keeping you under her gun shot is using physical coercion to modify your decisions.

[12] To account for supraindividual objects, and in general for supervenience, we need formal systems that help us deal with ontological levels. This is the main reason why we join to ATEL a first-order system that helps dealing with property change at a different ontology.

- Agents may *share* or not their mental representations and in particular their goals with others: $M, s' \models$ SHARE-GOAL $(\{x, y\}, \varphi)$ iff $M, s' \models \text{GOAL}_x\varphi \wedge \text{GOAL}_y\varphi$.
- Agents may depend on others (and vice versa) for achieving their goals (see also Castelfranchi et al., 1992). $M, s' \models$ DEP-GOAL$(\{x, y\}, \varphi)$ iff $M, s' \models \text{GOAL}_x\varphi \wedge \langle\langle y\rangle\rangle\Diamond\varphi \wedge \neg\langle\langle x\rangle\rangle\Diamond\varphi$. In words, agent x depends on y for φ because she has no strategy to achieve it, while y has one.

These dimensions are not incompatible but do not necessarily co-exist.[13]

Agents may perform one and the same action to achieve a shared goal, and may depend on one another for a goal they do not share. Obviously, agents' cohesion increases when both dimensions are positive.

Cohesion is therefore a function of shared-ness and interdependence. For Γ being a set of agents, $C(\Gamma) = f(S(\Gamma), I(\Gamma))$, let us define the functions S and I to $\{0, 1\}$, to mean absence of presence of property.

- $S(\Gamma) = 0$ and $I(\Gamma) = 0$ then $C(\Gamma) = 0$. This is the case with atomic agents, among which there is no social cohesion.
- $S(\Gamma) = 0$ and $I(\Gamma) = 1$, which means that there is a goal agents do not share but on which they are interdependent. This situation can be described with $M, s' \models \neg\text{GOAL}_x\varphi \wedge \neg\ \text{GOAL}_y\varphi \wedge \text{GOAL}_z\varphi \wedge \neg\langle\langle y\rangle\rangle\ \Diamond\varphi \wedge \neg\langle\langle x\rangle\rangle\Diamond\varphi \wedge \langle\langle z\rangle\rangle \circ \langle\langle x, y\rangle\rangle\Diamond\ \varphi$. If there is an agent z that has φ as a goal and is endowed with a strategy to get x and y to cooperate to achieve φ, then we have a case of orchestrated cooperation (for this notion, see Conte & Castelfranchi, 1995).
- $S(\Gamma) = 1$, $I(\Gamma) = 0$. $M, s' \models \text{GOAL}_x\varphi \wedge \text{GOAL}_y\varphi \wedge \neg\langle\langle x, y\rangle\rangle\Diamond\varphi$. Here, agents are not motivated to cooperate to achieve the shared goal.
- $S(\Gamma) = 1$, $I(\Gamma) = 1$. Now $M, s' \models \text{GOAL}_x\varphi \wedge \text{GOAL}_y\ \varphi \wedge \neg\langle\langle y\rangle\rangle\Diamond\ \varphi \wedge \neg\langle\langle x\rangle\rangle\Diamond\ \varphi \wedge \langle\langle x, y\rangle\rangle\Diamond\ \varphi$. Here, agents are interdependent in a shared goal; goal achievement depends on their full cooperation (see Fig. 1). In the following, we will not consider the lower left quarter of this figure, where there is neither objective interdependence nor shared representations among the agents. In describing the other three quarters, we will show the main differences between collective and shared actions.

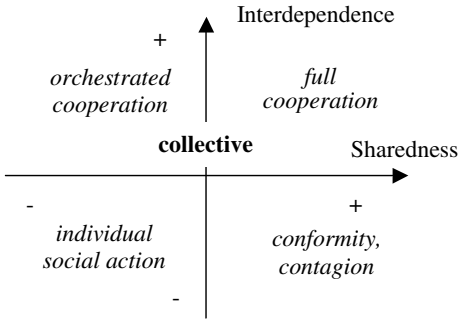

Fig. 1. Dimensions of multi-agent action.

6. Collective action

A couple of decades ago, a well-known philosophical debate occurred about the notion of collective action (Cohen & Levesque, 1990; Searle, 1990; Tuomela & Miller, 1988). In the Joint Persistent Goals' (JPG) model of collective action (Levesque, Cohen, & Nunes, 1990), JPG are mutually believed to be shared. To this view, Conte and Castelfranchi (1995) objected that collective action does not always imply mutual beliefs among performers. Hence, they proposed a more objective view, suggesting that an action done by a set of agents is collective when it is done to achieve a goal with regard to which the agents in that set are complementary:

> Collective action involves the elaboration of a plan to perform it, that is an action is collective when there is a plan to achieve φ by a set of agents Γ_i that are complementary for φ.

- *Complementary agents* (Conte & Castelfranchi, 1995) are all those agents that are necessary for a given world-state p to be realised. ($M, s' \models$ COMPLEMENTARY$(\{X\}),\varphi)$ iff $M, s' \models \langle\langle X\rangle\rangle\ \Diamond\ \varphi$ and $\forall x \in X$ not $M, s' \models \langle\langle X - x\rangle\rangle\ \Diamond\ \varphi$; that is, if X's elements are able to reach φ, then taken element x in X, X minus x is not able to reach φ anymore, or, otherwise stated, all X's elements are necessary although insufficient for φ to be reached.
- *Common Goal* (Conte & Castelfranchi, 1995): φ is common to a set of agents $\Gamma = \{x, y\}$ if it is shared and members of Γ are complementary with regard to it, where $M, s' \models$ COMMON-GOAL$(\{x, y\}, \varphi)$ iff $M, s' \models$ SHARE-GOAL$(\{x, y\}, \varphi) \wedge$ COMPLEMENTARY$(\{x, y\}, \varphi)$.

Formally then: $M, s' \models$ COLL-ACTION(x, y, φ) iff $\exists z \in \sum | M, s' \models \langle\langle x, y\rangle\rangle\Diamond\ \varphi \wedge\ \text{GOAL}_z\langle\langle x, y\rangle\rangle\Diamond\varphi \wedge\ \text{COMPLEMENTARY}\ (\{x, y\}, \varphi)$.

In natural language, x and y not only have a strategy to achieve a shared goal but they also have a goal to have such a strategy. By definition, collective action does not necessarily imply a common goal.[14]

[13] They are both semi-rigid properties, that is they are essential to some instancies but neither to all nor to none.

[14] Interesting is the notion of shared action, an action done by a set of agents when they achieve a shared goal: $M, s' \models$ SHARED ACTION(x, y, φ) iff $M, s' \models (\langle\langle x\rangle\rangle\Diamond\ \varphi \vee \langle\langle y\rangle\rangle\Diamond\ \varphi \vee \langle\langle x, y\rangle\rangle\ \Diamond\ \varphi) \wedge \text{GOAL}_y\varphi \wedge \text{GOAL}_x\varphi$.

Obviously, agents that share no goals and are not interdependent may engage in individual social action, both positive (for example exchange) and negative (aggression).

A set of complementary agents and the action they execute (the plan for that action) result in a collective agent, or briefly a collective, and we will denote it as C. The membership of the collective cannot simply be an algebraic relation. It is a relation of single agents that "belong to" the collective and their status of interpendency towards a goal.

Taken Φ the set of collective goals which are the goals for which agents that form a collective carry out a collective action, $\mathrm{GOAL}_C\,\Phi$. Collective agents are agents. Thus, every collective has a primitive goal. We call the set of primitives goals (that exist for (Theorem 1)) ψ^*: BELONG-TO($\{x, y\}$, C, p) = $\{x, y|$COLL-ACTION(x, y, p) and ($p \in \psi^*$ or MEANS-FOR (p, ψ^*)) Therefore, all the agents that carry out a collective action towards p, where p is either ψ^* or a means for it, belong if only transitorily to C.[15]

Of course, a number of crucial questions ought to be addressed here: what are the properties of the goals of the collective? Are they required to be coherent? How do they update? Of special interest, perhaps, is the question whom φ, the goal realised by C, belongs to. There are two cases.

First, φ is a goal of z's that does belong to the set of complementary agents. In such a case, C is what we have called orchestrated cooperation (see above, Fig. 1); for example, a plot organised by a criminal who exploits the actions of some innocent citizens that are complementary for her's goal. The action accomplished is collective, although the plot is not. Action is coordinated by the criminal, from which the executors are neither goal- nor plan-autonomous. Responsibility for the consequences can only be ascribed to who conceived the plot, although even the innocent citizen might be found responsible, if only to a lower degree. Consider that there may be orchestrated cooperation among collectives: sub-groups may inadvertently cooperate in plans that are designed by other groups. This is rather frequent among current criminal organisations, especially among terrorists, where it is extremely difficult to say who exploits whom for what.

In the second case, φ is shared by (a subset of) the complementary agents. Now, C is autonomous from external agents. Its internal autonomy is at stake here, and, as we shall see, this depends on other factors. It may be important to observe that although externally autonomous, C may undergo external influence, precisely as happens with autonomous individuals. In practice, this means that C must be endowed with some mechanisms for filtering external inputs, and possibly generating new goals as means for achieving old ones.

Complementary agents are likely to have specialised competencies. Often, in fulfilling their tasks agents are empowered (Jones & Sergot, 1996) by their roles in C. Empowerment may include some plan-autonomy. A given member of C that is charged with a given task may be autonomous as to how to achieve that task. Interestingly, one of such tasks may consist of filtering external inputs to C and take decisions as to whether or not to update its goals in the interest of C.

Suppose the collective gives the right to a member to count-as C in seeing to it that $\varphi : M, s' \models$ COUNT-AS(y, C, φ) iff $M, s' \models \mathrm{GOAL}_C\varphi$ and $\exists y \in \sum M, s' \models$ BELONGS-TO $(y, C, \varphi) \wedge (\langle\!\langle C \rangle\!\rangle \circ \Diamond\varphi \leftrightarrow \langle\!\langle C \rangle\!\rangle \circ \langle\!\langle y \rangle\!\rangle \Diamond\varphi)$.[16] That is, y counts as C for φ if and only if C has φ as goal and C has a strategy to ensure it if and only if C has a strategy to ensure that one of its member has it. In such a way, C becomes an agent, or better a special type of agent, directed by own goals and endowed with strategies (in this language, plans or pragmatic beliefs), both for updating goals and for achieving them, although these strategies must be executed by its members.

However, this representative, which counts as (Jones & Sergot, 1996) C with regard to this task, may fail to act in the interests of C for two reasons:

6.1. Error in decision-making

The member(s) counting as C may simply be mistaken and end up with the wrong decision: $M, s' \models \langle\!\langle C \rangle\!\rangle \circ \langle\!\langle y \rangle\!\rangle \Diamond \varphi$ and for all λ M, $\lambda[u > 1] \models \Box \neg\varphi$, because the delegated agent may not be able to carry out that task.

6.2. Cheating

If the representative is in turn an autonomous agent, it may be induced to generate and pursue a goal of its own, incompatible with the interests of C:[17] $M, s' \models \langle\!\langle C \rangle\!\rangle \circ \langle\!\langle y \rangle\!\rangle \Diamond \varphi$ and $M, s'' \models \neg\,\mathrm{Goal}_y \circ \varphi$ because the delegated agent may not want to carry out the task anymore. Worse, it may profit from its role to act in its own interests. Factors neutralising or reducing such a rise are the representative's accountability for the consequences of its decisions and the revocation of its mandate.

We can think that collectives can delegate to some members the updating of their own epistemic states. They become 'the mind' of the organization. For instance, a collective can ask some representatives to see to it that φ is a new goal of the collective. Such a new goal cannot be easily dropped: $M, s' \models \langle\!\langle C \rangle\!\rangle \circ \langle\!\langle y \rangle\!\rangle \circ [\![C]\!] \circ \mathrm{Goal}_C\varphi$.

As individual agents, also collectives have got mechanisms for deciding how to pursue hardwired goals (plan-autonomy)

[15] The relation MEANS-FOR is objective, that is agents belong even if they do not think so (but their actions realize the goals of the collective) and may not belong to the collective even if they think so (because their action do not actually belong to collective goals).

[16] Count-as in the sense of Jones and Sergot (1996). Although this notion is far more complex and it would need a much deeper account, it can be approximated by saying that the organization C has a strategy to make some of its members achieve one of its goals.

[17] By definition agents share some goals of the collective they are in. But we can imagine collectives generate new goals, and that these goals have strength, so they can be overlooked before stronger ones.

Table 1
Types of collective autonomy

	External aut.	Int. of collective	Int. of members
Terrorist cell	N	Y	N
Band of rogues	Y	N	Y
Firm	Y	Y	N

and filtering external inputs, thereby deciding which goals to have (goal-autonomy).

6.3. Limited plan-autonomy

Furthermore, as individual autonomy also collective autonomy is limited. Collectives are comparable as to their degree of autonomy (see Table 1). The larger, within the set of a collective's goals, the subset of goals that can be planned by that collective through institutional facts (cf. Searle, 1995), the more that collective is plan-autonomous.

Let us define a function plan-a§$\{1, 1\}(C)$[18] such that:

If $C = \{a\}$, that is if C is atomic, then plan-a§$\{1, 1\}$-$(C) = \#\Gamma$, where Γ is the set of goals about which an agent is plan autonomous. $\Gamma = \{p|P\text{-AUTONOMOUS } (a, \sum -C, p)\}$.
If C is a set of agents; that is $C = \{x_1, \ldots, x_n\}$, then plan-a§$\{1, 1\}(C) = \#(\Gamma' n)$ It is simply the cardinality of the set of goals they have and for which they are plan autonomous: $\Gamma' n = \{p|P\text{-AUTONOMOUS } (x_i, \sum -C, p)\}\}$.
If C is a collective then plan-a§$\{1, 1\}(C) = \#(\Gamma'' n)$, that is the cardinality of the set of goals they carry out as collective action, without being able to drop or generate them: $\Gamma'' n = \{p|P\text{-AUTONOMOUS } (C, \sum -C, p)\}\}$.

6.4. Limited goal-autonomy

The larger, within the set of a collective's goals, the subset of new goals that are accepted, modified or dropped by that collective through institutional facts, the more the collective is goal-autonomous. so goal-a§$\{1, 1\}(X)$:

If $C = \{a\}$, that is, if it is atomic, then goal-a§$\{1, 1\}$-$(C) = \#\ \Gamma$, where Γ is the set of goals about which an agent is goal autonomous: $\Gamma = \{p|G\text{-AUTONOMOUS } (a, \sum -C, p)\}$.
If $C = \{x_1, \ldots, x_n\}$, that is, it is a set of agents, then goal-a§$\{1, 1\}(C) = \#(\Gamma' n)$. This is simply the cardinality of the set of goals they have and for which they are goal autonomous: $\Gamma' n = \{p|G\text{-AUTONOMOUS } (xi, \sum -C, p)\}$.
If C is a collective then goal-a§$\{1, 1\}(C) = \#(\Gamma'' n)$, that is the cardinality of the set of goals the collective is able to generate: $\Gamma'' n = \{p|G\text{-AUTONOMOUS } (C, \sum -C, p)\}$.

[18] We denote functions with "§" and the couple $\{n, m\}$ where n is the arity of the function, m is the distinctive index. Moreover, relations are denoted with capital letters: $A\{1, 1\}$ is a unary relation, that is a property.

If we combine these two notions, we obtain that a collective is autonomous when it has got at least some level of plan-autonomy, i.e. when constitutive rules exist according to which it is possible to decide how to achieve existing goals. Instead, a collective will be goal-autonomous when, through institutional facts (constitutive rules), its goals can be modified. In practice, this means that the goals of an autonomous C cannot be modified offline, but only if they match C's criteria.

7. Supraindividual autonomy

From a certain point of view, participating in a collective action reduces autonomy. The head of a car convoy suffers some limitations in her decisions. As she is responsible (and accountable) before the other members of the convoy whom she is committed to (Conte & Paolucci, 2004), she is not allowed to stop at her will. The common goal, which limits the members' autonomy, is the filter of external input to C. C is granted external autonomy to the extent that the autonomy of its members is limited.

Here, however, the autonomy of the collective's members will be considered neither exclusively nor primarily (although we will examine to some extent the link between their autonomy and the autonomy of the supra-individual entity). The reader's attention here is drawn on the distinction between shared autonomy, which refers to the members, and collective autonomy, which requires that autonomy be extended to the supra-individual level. As supraindividual action, this includes shared and collective autonomy. However, whereas the former consists of a set of agents characterised by a comparable properties, only collective autonomy refers to the characterisation of a new entity, the Collective.

8. Collective autonomy

As individual autonomy, collective autonomy is a gradable notion. In short, we can say that given the set of all of a collective's goals, the larger the subset of goals upon which it takes decisions by means of constitutive rules, the more that collective is autonomous. This is true for plan-autonomy, and a fortiori, it is true for goal-autonomy.

When is C autonomous? To answer this question, different aspects of collective autonomy must be analysed:

- Autonomy to do what? This question has received a certain attention within the literature on autonomy in delegation (Castelfranchi & Falcone, 2004), teamwork (Martin & Barber, 1996; Sierhuis, Bradshaw, van Hoof, Jeffers, & Uszok, 2003; Tambe, 1997), and organizations (Mullarkey, Jackson, & Parker, 1995; Spriggs, Jackson, & Parker, 2000; Weber, 1999) and we will not dwell on it here.

- Autonomy of whom – C or its members – from what – internal or external entities?

9. Autonomy from what?

The issue of what concretely means for a collective to be autonomous, from whom and to whom autonomy is attributed, has been relatively overlooked in MAS and the organizational literature (for exceptions, see Marion, 1999; Raelin, 1989; Sewell, 1998; Willmott, 1993).

As any other entity, C can be plan-autonomous or goal-autonomous. In the first case, C can achieve goals that can be manipulated, modified or dropped by some entities without being accountable for that. C is also goal-autonomous when there is no entity, either internal or external to $\sum$, which can change its goals offline. In practice, this means that a lot of collective entities are only plan-autonomous. But not all of them: State, Parliament, etc. are often informally said to have a "life of their own", to be "ends in themselves", etc. Our proposal is to account for this intuitive characterisation in terms of goal-autonomy. However, let us distinguish internal from external autonomy and examine the relationships between the system's autonomy and that of its members (see Table 1).

10. Outer or external autonomy (ex-autonomy)

Much like individual autonomy, there is no z external to Y that can change C's goal(s) offline.

External autonomy means that C's goals may be modified only based on a decision of C's (or of its charged members).

A corollary of this principle is that C is ex-autonomous before entity E external to C, if for any state of the world q such that E wants C to have q as a goal, C will do so if there is at least another, older goal p of C's that C (its charged members) believes q to be a means for. $M, s' \models$ EX-AUTONOMOUS (C, E, q) iff for all $\lambda, M, \lambda[1] \models \text{Goal}_C q$ and $\exists p \neq q \in \Pi, \exists L \in \sum |M, s' \models \text{Goal}_C p \wedge$ BELONG-TO $(L, C, p) \wedge$ COUNT-AS $(L, C, q) \wedge \text{Bel}_L$MEANS-FOR$(q, p) \wedge \text{Goal}_C \text{Goal}_C q \wedge [\![E]\!] \circ \text{Goal}_C q$. In words, a collective C is ex-autonomous towards E regarding the *sw* q, iff C gets to have a goal q thanks to L, which counts as C, believing that q is a means for goal p that already is a goal of C's.[19] In the goal generation of collectives, representatives can change goals without necessarily adopting these.

Of course, C must be influenced from the external world, otherwise it would be a self-referential system, poorly modifiable and hardly adaptable. We know that thanks to the principle of limited autonomy, a system can be autonomous and at the same time undergo external influence. In particular, cognitive influencing is a goal-directed action aimed to modify the representations (goals and beliefs) of the target entity in order to modify its behaviour (Conte & Castelfranchi, 1995).

When trying to apply it to concrete institutions, one realises that the notion of ex-autonomy is relational. For example, whereas public institutions are not ex-autonomous with regard to the State, they are such with regard to private entities. On the other hand, private organisations as well as non-governmental institutions are entirely ex-autonomous. Interestingly, modern institutional democracies are founded on the principle of division and balance of power. Hence, judiciary, legislative and executive powers are equally ex-autonomous.

11. Inner autonomy: inn-autonomy

A collective entity is internally autonomous when it is (relatively) independent from the goals of its members. In particular a distributed collective is internally autonomous when a subset of its members has the specialised task to filter inputs and requests from inside. The stronger this filter, the more internally autonomous the collective is.

More explicitly,

> C is inn-autonomous if for any member x of C and any world-state q such that x wants C to have q as a goal, there is at least an older goal of C's for which C (its representatives) believes the new goals to be means: $M, s' \models$ INN-AUTONOMOUS (C, x, q) iff for all $\lambda, M, \lambda[1] \models \text{Goal}_C q \wedge$ BELONG-TO$(x, C, q) \wedge [\![x]\!] \circ \text{Goal}_C q$ and $\exists p \neq q \in \Pi$ and $\exists L \in \sum |M, s' \models \text{Goal}_C p \wedge$ BELONG-TO$(L, C, p) \wedge$ COUNT-AS$(L, C, q) \wedge \text{Bel}_L$ MEANS-FOR$(q, p) \wedge \text{Goal}_C \text{Goal}_C q$.

12. Autonomy of whom?

Whom is internal autonomy predicated of, the system or its members? These two aspects are strictly interdependent, as we shall see, but we should keep them conceptually distinct. In particular, participation in C has costs, as it reduces the members' autonomy: they cannot drop their commitments, must accept collective decisions, etc. Now the thesis that will be argued here is that internal autonomy of C from members implies that C cannot be bent to personal or particular interests. C is endowed with power and mechanisms to decide upon internal inputs coming from subsystems or individuals.

As individuals, collectives are limited autonomous systems; hence, an entirely independent collective is probably non-existent. Both ex-autonomy and inn-autonomy are relational and gradable notions, which must be considered from two complementary points of view: that of the system and that of its members.

From their side, sub-systems are granted external autonomy: no-one outside the system can change their goals without acting upon C. On the other hand, they

[19] The collective assigns to some members the power (empowers) to decide wether a certain goal has to be pursued by the collective itself. The representatives can be even equal to the members of the collective itself (assembly) or to one (dictatorship, charismatic leadership, etc.).

can be more or less autonomous from C. Of course, their autonomy implies a limited commitment to C, what is likely to lead to C's desegregation.

From its side, C may be inn-autonomous with regard to some of its members (employees) but not to others, i.e. the share-owners working within the enterprise. In this case, the higher the share, the more C depends on (some of) its members. Furthermore, C (e.g., a firm) can be autonomous from external but not from internal entities. Another (e.g., a terrorist cell) may enjoy a poor external autonomy, but must be autonomous from its members. Finally, there are systems which are neither externally nor internally autonomous, and this is the case of a claque. On the other hand, a band of rogues has got both external and internal autonomy.

13. The virtuous (or vicious) circle of collective autonomy

The following claims seem to derive from the analysis unfolded so far: The less inn-autonomous its members, the more inn-autonomous C. We state inn-autonomy of a single agent is his/her autonomy from the collective she/he is in.

Theorem 2. *The more an organization C is inn-autonomous, the less its members are. And vice versa.*

Proof. Inn-autonomy augmentation can be represented as a relation between a state of the world and its immediate successor, comparing the collective autonomy and the individual ones.

Being C a collective and given BELONG-TO(x_i, C, Γ) $\forall\, s \in Q$, s(inn-aut(C)) $> s_{+1}$(inn-aut(C, $U\, x_i$)) $\rightarrow s$(aut($U\, x_i$)) $< s_{+1}$(aut($U\ x_i$)). □

Definition 1. inn-aut(C, $U\, x_i$) = #{p|INN-AUTONOMOUS (C, $U\, x_i$, p)}.

Definition 2. aut($U\ x_i$) = #{p|AUTONOMOUS(x_i, $\sum -$ $U\ x_i$, p)}.

But $\forall p \in \Pi$, INN-AUTONOMOUS $(C, U\, x_i, p) \rightarrow \neg$ AUTONOMOUS ($U\ x_i$, C, Goal$_C p$) because, from inn-autonomy of C it follows that its members cannot avoid that C cannot avoid to have p has a goal, once generated.

Given a set of goals Γ, if C_x and C_y are two collectives, then if inn-aut(C_x, $U\, x_i$, Γ) > inn-aut (C_y, $U\, x_i$, Γ) then $\exists\ p \in \Gamma|$ AUTONOMOUS ($U\, x_i$, C_x, p) and not AUTONOMOUS ($U\ y_i$, C_y, p). Therefore inn-aut(C_x, $U\ x_i$) > inn-aut(C_y, $U\ y_i$) $\rightarrow$ aut($x_i, \sum$) < aut($y_i, \sum$).

In words. If the internal autonomy of a collective C_x towards its members is major than the internal autonomy of an other collective C_y towards the same set of agents, then there exists a p such that the members of C_y are autonomous towards the collective whereas the members of C_z are not. That is the autonomy of the members of C_x towards is minor than the autonomy of members of C_y, which our theorem is a particular instance of.

The more C is inn-autonomous, the less it is bent to personal or private interest: if s (inn-aut(C, $U\ x_i$)) > s_{+1}(inn-aut(C)) $\rightarrow s$ (aut($U\ x_i$)) < s_{+1}(aut($U\ x_i$)) (Theorem 2), means that internal autonomy growth decreases the cardinality of goals agents can autonomously pursue (generate) or drop, remaining inside the organization they belong to.

Indeed, the notion of collective autonomy lends itself to characterise and classify institutional systems in terms of the relative degrees of freedom that their members are granted.

14. Why bother with collective autonomy?

So far we provided a conceptual analysis and some theoretical statements, to be either proved or tested. The question is: what is the use of this analysis? Which advances are we allowed thanks to it?

Potential contributions of the present work range from the study and management of organisations (networking) to AT and MAS.

Current work in organizations theory (see Prietula, Carley, & Gasser, 1998) explores the mutual impact of individual cognition and structural and procedural aspects of organizations. More radically, in the present work agent properties are extended to the level of the organisations in order to model them and provide a theoretical background for the formulation of hypotheses concerning their evolution. Two operational sets of hypotheses about the dynamics of organizations arise from the present analysis: (a) the dimensions of interdependence and shared-ness ought to be further explored to put forward testable hypotheses about systems' cohesion and, possibly, their trade-off between stability and efficiency; (b) external and internal autonomy of collectives provide criteria for systems' comparison and classification: both affect systems' stability and preservation, but whereas the former is a affect their preservation, but in different ways. As sources of destabilization come both from the outside and the inside, inn-autonomy is more directly responsible and a better predictor of collectives' persistence and good health: collectives "standing on their own", may have argyle feet. High ex-autonomy inevitable leads to an increased exposure of the collective to external pressures and potential (external or internal) attacks. Consequently, ex-autonomy is not a good controller of vulnerability, nor a good predictor of robustness. Inn-autonomy, instead, is a stronger predictor of reduced vulnerability and effective stability of collectives. It also can highlight conditions for their evolution/involution: a promising direction of study that unfolds from the model here presented, beyond the scope of the present collection, is the impact of internal autonomy on political systems, and especially on the likelihood of authoritarian turns. What is the difference between coercive, authoritarian and totalitarian systems: how and to what extent can the present analysis contribute to highlight the conceptual network in which these notions are plunged, and define predictors of the corresponding empiri-

cal phenomena? A further related question concerns the mutual impact of the two dimensions of collective autonomy: is ex-autonomy leading to inn-autonomy or the other way around? Or are they inconsequential on each other? This is a subject for future investigation, but since internal upholds are intrinsically ominous, inn-autonomous systems are more likely to prevent and resist disintegration and/or subversion.

The second direction of application of the proposed model, agent and multi-agent systems, stems from further investigation concerning the interplay between internal autonomy of the system and autonomy of its members. Such a trade-off highlight important questions for the design and management of intelligent software agents in cooperative and competitive interaction: in particular, is there a critical, equilibrium depending on circumstances and conditions that might be aimed at, between the degree of autonomy of members and that of the system? Plausibly, members' autonomy interferes with their performance, especially with their flexibility and efficiency. Hence, to what extent members can be granted autonomy, without endangering the integrity and efficacy of the system? On the other hand, how does the system autonomy (both internal and external) reverberate on expected efficacy of the system, hence on members respecting their commitment, and finally on the system's effective efficacy?

References

Castelfranchi, C., & Falcone, R. (2004). Founding autonomy: the dialectics between (social) environment and agent's architecture and powers. *Lecture Notes on Artificial Intelligence, 2969*, 40–54.

Castelfranchi, C., Miceli, M., & Cesta, A. (1992). Dependence relations among autonomous agents. In E. Werner & Y. Demazeau (Eds.), *Decentralized AI*. Amsterdam: Elsevier Science Publishers.

Cohen, P. R., & Levesque, H. J. (1990). Intention is choice with commitment. *Artificial Intelligence, 42*, 213–261.

Conte, R., & Castelfranchi, C. (1995). Cognitive and social action. London: UCL Press.

Conte, R., & Paolucci, M. (2004). Responsibility for societies of agents. *Journal of Artificial Societies and Social Simulation, 7*(4). Available from http://jasss.soc.surrey.ac.uk/7/4/3.html.

Franklin, S., & Graesser, A. (1996). Is it an agent or just a program? A taxonomy for autonomous agents. In *Proceedings of the third international workshop on agent theories, architectures and languages*. Berlin: Springer.

Guarino, N., & Welty, C. (2000). A Formal ontology of properties. In R. Dieng (Ed.), *Proceedings of the 12th international conference on knowledge engineering and knowledge management. Lecture notes on computer science*. Berlin: Springer.

Guarino, N., & Welty, C. (2002). Identity and subsumption. In R. Green, C. A. Bean, & S. Hyon Myaeng (Eds.), *The semantics of relationships: an interdisciplinary perspective* (pp. 111–126). Kluwer.

Hexmoor, H. (2004). A cognitive model of situated autonomy. csce.uark.edu/~hexmoor/CV/PUBLICATIONS/CONFERENCES/PRICAI-00/PRICAI00-4.doc.

Jones, A., & Sergot, M. (1996). A formal characterisation of institutionalised power. *Journal of the IGPL, 3*, 427–443.

Levesque, H. J., Cohen, P. R., & Nunes, J. H. T. (1990). On acting together. In *Proceedings of the eighth national conference on artificial intelligence (AAAI-90)* (pp. 94–99), Boston, MA.

Marion, R. (1999). The edge of organization: Chaos and complexity theories of formal social systems. Thousand Oaks, CA: Sage.

Martin, C., & Barber, K. (1996). Multiple, simultaneous autonomy levels for agent-based systems. In *Proceedings of the fourth international conference on control, automation, robotics, and vision*.

Mullarkey, S. L., Jackson, P. R., & Parker, S. K. (1995). Employee reactions to JIT manufacturing practices: a two-phase investigation. *International Journal of Operations and Production Management, 15*(11), 62–79, 18.

Prietula, M., Carley, K., & Gasser, L. (1998). *Simulating Organizations: Computational Models of Institutions and Groups*. The MIT Press.

Raelin, J. (1989). The anatomy of autonomy: managing professionals. *Academy of Management Executive, 3*(3), 216–228.

Searle, J. (1990). Is the brain's mind a computer program? *Scientific American, 262*(1), 20–25.

Searle, J. (1995). *The construction of social reality*. New York: Free Press.

Sierhuis, M., Bradshaw, J. M., Acquisti A., van Hoof, R., Jeffers, R., & Uszok, A. (2003). Human-agent teamwork and adjustable autonomy in practice. In *Proceeding of the 7th international symposium on artificial intelligence, robotics and automation in space: I-SAIRAS 2003*, NARA, Japan, May 19–23, 2003.

Sewell, G. (1998). The discipline of teams: the control of team-based industrial work through electronic and peer surveillance. *Administrative science quarterly*, June, Special Issue on "Critical Perspectives on Organizational Control".

Spriggs, C., Jackson, P. R., & Parker, S. K. (2000). Production teamworking: the importance of interdependence and autonomy for employee strain and satisfaction. *Human Relations, 53*(11), 1519–1543.

Tambe, M. (1997). Towards flexible teamwork. *Journal of Artificial Intelligence Research*, *7*, 83–124.

Tuomela, R., & Miller, K. (1988). We intentions. *Philosophical Studies, 53*, 367–389.

Van der Hoek, W., & Wooldridge, M. (2003). Cooperation, knowledge and time: alternating-time temporal epistemic logic and its applications. *Studia Logica, 75*, 125–157.

Weber, W. G. (1999). Kollektive Handlungsregulation, kooperative Handlungsbereitschaften und gemeinsame Vergegenständlichungen in industriellen Arbeitsgruppen. *Zeitschrift für Arbeits- und Organisationspsychologie*, *43*. Jg., Heft 4.

Willmott, H. (1993). Strength is ignorance; slavery is freedom: managing culture in modern organizations. *Journal of Management Studies, 30*(4), 515–552.

Further reading

Arednt, H. (1951). The origins of totalitarism. New York: Harcourt.

Fest, J. (1999). Hitler. Eine biographie. Berlin: Rowohlt-Verlage.

Fest, J. (2000). Speer. Einer biographie. Berlin: Rowohlt-Verlage.

Fest, J. (2005). Der Untergang, Hitler und das Ende des Dritten Reiches. Eine historische Stizze. Berlin: Rowohlt-Verlage.

Held, D. (1987). Models of democracy. Cambridge: Polity Press.

Held, D. (Ed.). (1991). *Political theory today*. Cambridge: Polity Press.

Held, D. (1995a). *Democracy and the global order. From the modern state to Cosmopolitan governance*. Cambridge: Polity Press.

Held, D. (1995b). Democracy and the new international order. In D. Held & D. Archibugi (Eds.), *Cosmopolitan democracy. An agenda for a new world order*. Cambridge: Polity Press.