Rosaria Conte and Mario Paolucci (2004)

Responsibility for Societies of Agents

Journal of Artificial Societies and Social Simulation vol. 7, no. 4
<http://jasss.soc.surrey.ac.uk/7/4/3.html>

To cite articles published in the *Journal of Artificial Societies and Social Simulation*, reference the above information and include paragraph numbers if necessary

Received: 12-Dec-2003 Accepted: 31-May-2004 Published: 31-Oct-2004

Abstract

This paper presents a pre-formal social cognitive model of social responsibility as implying the deliberative capacity of the bearer but not necessarily her decision to act or not. Also, responsibility is defined as an objective property of agents, which they cannot remit at their will. Two specific aspects are analysed: (a) the action of "counting upon" given agents as responsible entities, and (b) the consequent property of accountability: responsibility allows to identify the locus of accountability, that is, which agents are accountable for which events and to what extent. Agents responsible for certain events, and upon which others count, are asked to account or respond for these events. Two types of responsibility are distinguished and their commonalities pointed out: (a) a primary form of responsibility, which is a consequence of mere deliberative power, and (b) a task-based form, which is a consequence of task commitment. Primary responsibility is a relation between deliberative agents and social harms, whether these are intended and believed or not, and whether they are actually caused by the agent or not. The boundaries of responsibility will be investigated, and the conceptual links of responsibility with obligation and guilt will be examined. Task-based responsibility implies task- or role-commitment. Furthermore, individual Vs. shared Vs. collective responsibility are distinguished. Considerations about the potential benefits and utility of the analysis proposed for in the field of e-governance are highlighted. Concluding remarks and ideas for future works are discussed in the final section.

Keywords:
Responsibility, Agents, Cognitive Modeling, E-Governance, Organisation Theory

Introduction

1.1 There is much rumour about an increasing impulse to 'responsibilise' citizens and corporations, to take advantage of responsible agency for the sake of governance aims, i.e. social-cultural

regulation at a distance rather than through the imposition of norms (Lacey 2001).

1.2 Responsibility is of analogous utility also in the context of information and communication technology, with special emphasis on agent societies (Mamdani and Pitt 2000). The need of *e*-regulation, *e*-government, and *e*-institutions is in the agenda of scientists and designers who care about trustability and adaptation of technologies to the users (see Ören 2000). On the other hand, computer ethics[1] has reached an important point, influencing policy formulation, computing practice and computer application.

1.3 Responsibility has been investigated in several fields of science: from social and cognitive psychology (Heider 1958), to philosophy of law and deontic philosophy and logic (Hart and Honoré 1985; but see also Jones and Sergot 1996.). Even in AI and MAS, attempts to formalize this notion exist (Jennings 1992; Jennings and Mamdani 1992; Santos et al. 1997; Norman and Reed 2001; 2002).

1.4 This paper presents a different approach to this issue. Unlike the social psychological approach, it is not interested only in the social "attribution" of responsibility. Unlike the deontic approach, it is not focused on legal or institutional responsibility, but on social responsibility and its micro-foundations. Unlike previous work in AI and MAS, it defines responsibility as an objective and emergent property of agents. In our approach, agents are responsible for social harms, whether they perceive and cause them or not. Responsibility, then, is an attribute of the agent, rather than a state represented in her mind. However, it presupposes a deliberative architecture of the mind.

1.5 The view proposed in this paper can be resumed in the following statements:

- Responsibility implies deliberative capacity (free will, etc.): only autonomous and deliberative agents can be responsible for given events, both negative and positive. It does not automatically imply (nor excludes) a decision to act or not act.
- Responsibility is attributed on the grounds of agents' powers: this does not prevent agents from "feeling responsible" about events, but we will not concentrate upon this issue.
- There are different types of responsibility. We will distinguish a primary form of responsibility, which is a consequence of mere deliberative power, from a task-based form, which is a consequence of role and task assignment.
- Responsibility is the objective grounds upon which two important phenomena are engrained: (a) the mental state of "counting upon" agents as responsible entities; (b) the consequent property of accountability: responsibility allows to identify the locus of accountability, that is, which agents are accountable for which events and to what extent. Agents responsible for certain events, and upon which others count upon, are asked to account or respond for these events. But what kind of mental state is to count upon someone? And what is implied by accountability?
- Responsibility is an objective property, one which agents cannot remit at their will. This allows accountability.

1.6 In the following, after a brief review of some relevant literature on the subject matter, the two main notions will be distinguished, primary and task-based responsibility. Primary responsibility will be proposed as a relation between a deliberative agent and a subset of the world states, whether these are intended and believed or not, and whether these are actually caused by the agent or not. The boundaries of responsibility will be investigated, and the conceptual links of responsibility with obligation and guilt will be examined. Task-based responsibility will be argued to imply task- or role-commitment. In section 4, individual Vs shared Vs collective responsibility will be distinguished. In section 5, considerations about the potential benefits and utility of the analysis proposed for the agent systems field will be highlighted. Concluding remarks and ideas for future works will be discussed in the final

section.

Related Work

2.1 The notion of responsibility has an important philosophical, legal, and social psychological tradition. I will not re-examine here this immense literature, but will only point out the open questions that are particularly relevant for the present context. Therefore, existing views will not be critically examined under the aspect of responsibility *attribution,* but rather under a general problem of applied ethics: what is responsibility? Is it a specific notion, to be kept distinct from guilt, on one hand, and from causation, on the other? Why bother with responsibility, what is the utility of this notion in a changing technological, ethical, organisational and legal environment? What type of agency is implied by responsibility? More explicitly, which properties are agents required to possess in order to be responsible?

2.2 Analytical treatment of responsibility goes back to Aristotle, according to which "people should be held responsible for the outcomes of exactly those choices that were free and unaffected by circumstances." (Hild and Voorhoeve 2001). For the Greek philosopher, responsibility holders are decision-makers endowed with the capacity to foresee consequences of action (or inaction) and choose accordingly.

2.3 While this notion points to a general property of deciding agents, later Greek or Latin thinkers pointed to a notion of responsibility that is closer to what recent philosophers and sociologists have called *role* (French and Raven 1959), *de jure* (Hart and Honoré 1985) or *rule-following* responsibility (Zsolnai forthcoming). Plato, in the *Republic*, distinguished different roles within the ideal hierarchical state, each associated with the corresponding *responsibility*. Analogously, Cicero in his *De officiis,* provided a well-developed account of role responsibilities (in Latin, the word *officium* means duty or responsibility) (for a good review, see Mitcham and von Schomberg 2000). Later, philosophers of law (Hart and Honoré 1985) have distinguished *de facto* and *de jure* responsibility, the former being inherent to human beings, the latter being defined by legal terms or contractual definitions.

2.4 What is the nexus, if any, between the two phenomena, and what about a general definition of responsibility?

2.5 To make a complicated issue even more puzzling, other dimensions of responsibility appear to be intertwined with this fundamental dichotomy, e.g. prospective and retrospective responsibility, and therefore *ex post* or *ante hoc* attribution (see again Hart and Honoré 1985; Zsolnai forthcoming). Furthermore, responsibility appears to bring about a set of related concepts, such as *accountability* and *liability* (Seeger 2001). At a more abstract level, other notions are called into question, such as *guilt*, *obligation*, and *causation* (Seeger 2001 and many others). Another question particularly dear to moral philosophers (Sartre 1984; Heidegger 1977) and to political scientists (Arendt 1969; Lenk 1997), is the relationship between responsibility and *freedom*. Since Aristotle, scholars have long debated around an ambiguity they found in the Greek philosopher, concerning the function of blame. Is blame appropriate because the agent *deserved* it, or is it appropriate because it *leads* agents to do better in the future? In causal determinism (and in its scientific and teleological variants), many found an implicit threat to moral responsibility: anything which has been caused by sufficient antecedent conditions, cannot be blamed. It was not until the Stoic philosophy (third century BC), that a fundamental distinction was introduced between determinism and *fatalism*. If causal antecedents include such things as deliberation, choice, and action, they will contribute to determine effects.

2.6 The debate whether causal determinism and moral responsibility are compatible or not has puzzled philosophers for ages. For those who accept the Stoic interpretation, causal

determinism is not necessarily incompatible with moral (check the *Stanford Encyclopaedia of Philosophy* under 'moral responsibility'), whereas for others, the two views are incompatible. In general, the former view is associated to a merit-based conception of responsibility (see Strawson 1974), i.e. to the idea that blame is appropriate because agents deserve it. The latter view is associated with the consequentialist conception, in which blame serves to improve future action.

2.7 Finally, the possibility to attribute responsibility to higher-level entities (*collective responsibility*) or even to aggregates and groups of agents (*shared responsibility*) is not always solved in the most convincing way (Ehrenberg 1999; Lane 2004; Silver 2002). This dimension of the issue is independent of the preceding dilemma, i.e. consequentialist *versus* merit-based responsibility. Whether responsibility is defined in terms of action preconditions (merit-based) or in terms of unforeseen consequences (consequential) (Strawson 1974), its sharing always presupposes some level or degree of *joint action*. This, however, is too strong for moral and political philosophers willing to extend the range of co-responsibility far beyond the performance of common activity. For some (Moore 1998) if responsibility is grounded in human agency, it can be applied exclusively to individual human personhood.

2.8 Consequentialist authors argue that responsibility does not have to be congruent with *personal* control (cf Hart and Honoré 1985, Fleurbaey 1995, Scanlon 1998), since *post hoc* assessment not always allows to ascertain to what extent and which effects could have been foreseen. *Rule-following* responsibility is easier to assess: agents hold responsibility for executing a given set of rules associated to the roles they accepted to play, in the sense theorised by Plato and Cicero.

2.9 Such a solution is still of partial utility. It leaves open the question as to the effects of roles: is it appropriate to put the blame on Nazi officers? Is co-responsibility still a meaningful notion at all? The solution to this problem in terms of "ownership of group actions" (Silver 2002) seems rather too metaphorical.

2.10 In short, the notion of responsibility centred on agency and free-choice raised a lot of problems in the last two thousands years:

- Does responsibility imply a *decision* to act/not act? If so,
- What relationship does it hold with *obligation* and *causation*?
- Related to the preceding one, what is the difference between responsibility and *guilt*?
- What about *shared* responsibility? Are we happy enough with a co-responsibility shared by participants in joint decisions?
- What about *collective* responsibility? Can we attribute the properties of human agency to the level of collective entities?
- What about *de facto* responsibility? How can we assess, post hoc, predictability of effects of such decisions?
- What is the relationship between *de facto* and *de jure* responsibility? Are they distinct phenomena? Can we achieve a general theory of responsibility, which allows us to account for the interaction between these two notions, as is probably necessary to handle such complex moral and legal situations as the case of Nazi officers?
- What is the relationship between responsibility, on one hand, and *accountability*, on the other?

2.11 This paper presents a pre-formal theoretical analysis of the notion of responsibility as a necessary property of autonomous deliberative agents in multi-agent contexts, which *does not imply a decision to act*. As will be shown through the paper, this fundamental divorce allows to account for the preceding questions, or at least paves the way for the solution to most of them.

2.12 In contrast with the decision-based, free will-based and subjective view of responsibility, a conceptualisation of responsibility as an objective, emergent property, namely a special type of *power*, of intelligent autonomous agents will be proposed. As will argued in the rest of the paper, this solution allows

- To acknowledge responsibility for events which do not proceed from any decision of the responsibility holder
- To then distinguish responsibility from both causation and obligation, on one hand, and from guilt on the other, which seems instead to imply some sort of decision. Whereas responsibility is a power, guilt implies some decision.
- To account for shared and collective responsibility without assuming participation in joint decision; this extends the range of co-responsibility beyond joint action/inaction, for the sake of the moral and political thinkers à la Sartre.
- To account for the interaction of *de facto* and *de jure*, natural and role-based responsibility: both refer to a specific power of intelligent agents, which does not imply decisions to act. In this sense, a reliable rule-follower is still responsible for given effects if she is endowed with the responsibility *power*.

Responsibility as an Objective and Emergent Property

3.1 In the present paper, an *objective* notion of responsibility is proposed, in the sense that it can be assessed from an external observer. Deliberative agents are held responsible for the harming effects they have the *power* to prevent or reduce, although they did not actually intend nor caused them.

3.2 Furthermore, responsibility is an *emergent* property of agents. A deliberative agent can be attributed a further property, that of being responsible for the world states which it has the *power* to prevent. In part, the present analysis of responsibility accounts for the intuition (see also Lacey 2001) that responsibility is specified according to the moral or normative codes in force at a given time in a given group. Nonetheless, a core notion of responsibility as an objective property of intelligent social agents is probably shared by different societies in different historical and temporal conditions.

3.3 To clarify this core notion, we will distinguish a fundamental or *primary* form of responsibility (what many authors call "*de facto*" responsibility) from a *task-based* responsibility (closer to the notion of role-based or rule-following responsibility), which is a conceptual evolution of the former and cannot be understood without it. Both share a fundamentally *objective* character, in the sense previously defined; in addition, task-based responsibility implies that the responsible agent is aware and willing to accept a given role.

3.4 Some warnings about the present analysis are necessary. First, we do not refer to specific domains of responsibility, e.g. legal or moral, as some authors (Hart and Honoré) do, although in the following analysis examples drawn from these domains abound. In our view, responsibility does not imply the violation of norms or moral standards (see Shaver and Drown 1986). Nonetheless, we believe that responsibility has specific advantages at the society level (to redistribute the costs of repair and of prevention of social harm, as was pointed out by some philosophers, see Feinberg 1970). In particular, what will be examined here is the global effect of agents' *accountability*: a responsible agent is asked to respond for the effects it is responsible for and bear their consequences. What does this mean? What types of consequences is one expected to bear?

3.5 An urgent aspect of responsibility lies in its quantification, i.e. the factors contributing to increase or decrease responsibility. We will not focus on this important issue in this paper, although we will propose some dimensions for quantification throughout the paper. This aspect could be a relevant objective for future studies.

3.6 Some authors (Shaver and Drown 1986) state that there is no common notion of responsibility but only a set of dimensions along which to assess it. Conversely, we believe that a general theory of responsibility can be formulated. The analysis presented in this paper is certainly incomplete, but it is aimed to highlight some fundamental ingredients for such a theory.

Primary Responsibility

3.7 In this section, we will address the notion of being responsible for some world state.

3.8 What kind of property is this? Following Aristotle, we should say that a deliberative agent is responsible for any state of the world that is consequent to its (in)actions. From this point of view, the notion of responsibility collapses on the notion of decision.

3.9 Here, instead, a rather different notion is proposed: an agent is responsible for something when it *could* have avoided it, that is, when it has the power (whether internal or external, endogenous and exogenous) to avoid it.
Def: Responsibility

Agent *x* is responsible for the world state *s*, when *x* can prevent *s*

Of course, *s* can be either potential or actual. An agent that has the power to prevent a given event, and does (not) exert it has a responsibility on that event once it has occurred. Why such an emphasis on the power to prevent *s*, rather than on the power to restore it? This will become clearer later on in the paper. Intuitively, agents are considered responsible, and often are asked to account for, events which they let occur, and not only for events which they have directly caused. On the other hand, agents may cause events that they cannot be held responsible for, since they had no power to avoid them. In Cohen and Levesque (1990) terms,

3.10 Note also that a subtle distinction could be made here between the believed and effective avoidance capability. An agent could believe that *s* is unavoidable (true in all its belief accessible worlds), but some of these beliefs could be false; the same could happen in the other direction, i.e., an agents could believe itself responsible for an event that is instead unavoidable given the real world. We will not consider these distinctions in the rest of the article; all our definition can be used both at for believed and the real world.

3.11 This definition is still independent of any collective. But we are interested here mainly is an analysis of *social* responsibility: we propose a definition where agents are socially responsible when they have the power to avoid a social harm (whether this is caused by their (in)actions or not). But what is a social harm? The following considerations suggest a preliminary definition:

- The victim should not be necessarily meant as an individual agent, but also as a social system. Indeed, agents may have responsibilities for public goods, institutions, etc.
- Agents are responsible for *objective* harms. A victim's perception is not necessary: gossip can lead to a loss of reputation of the victim, whether this knows it or not. In this case, *x* (in this example, the gossiper) is responsible before external observers, and sometimes before specified authorities (be their institutional, like a justice court, or non-institutional, like the public opinion).
- Agents are socially responsible for *socially relevant* harms. The victim's perception is not sufficient either. Successful agents may cause others to suffer from their achievements, but they are not *socially* responsible for these pains. Suffering from envy is not socially relevant. Although the notion of socially relevant harm is a rather complex one, we will consider here one [2] dimension of it, namely the victim's *loss* of power, means and resources (life, physical integrity, resources, reputation, liberty, dignity, etc.). In possible worlds semantic, this could be evaluated by some decreasing of the goal accessible futures, or by evaluating possible future with some utility function for the agents involved.

Def: Social Harm

World state *s* is a social harm, when

- *s* implies a loss of a given agent (possibly also an institutional agent, see Carmo and Pacheco 2001) or set of agents *y*; that is, *s* a world state which reduces their power to achieve their goals
- even independent of *y*'s beliefs

3.12 With the definition for social harm, we can state the corresponding one for social responsibility.

Def: Primary Social Responsibility (base)

Agent *x* is responsible for the world state *s*

- when *x* can prevent or reduce s, and
- *s* is a social harm (at the expense of *y*)

3.13 The definitions given are summarised in figure 1, where social responsibility is shown to specialize responsibility depending upon social harm.

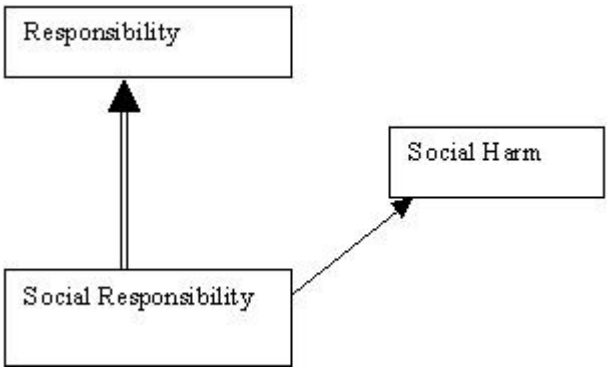

Figure 1. Responsibility generalises (double line) social responsibility and is defined upon (full line) social harm

However, this definition does not take into account the motivation of *x*. If *s* is an actual event, two possibilities occur: either *x* did not exert its power or this was not sufficient to prevent *s*. Only in the former case, *x* can be blamed and is accountable for *s*. Let us discuss this aspect of responsibility.

Power

3.14 In the present analysis, *x* is not considered responsible for the harms it has caused (*negative power*), but for those it might have *avoided (aversive power*). Agents are responsible for the harming consequences of their or others' actions if and only if they had the power to avoid these consequences: bystanders who do not intervene in social emergencies (rapes, burglaries, etc.) usually are held (co-) responsible for the victims' injuries.

Def: Negative Power

- *x* can bring about *p*
- for a given agent *y* and a given goal *q* of *y*, *p* implies not *q*.

Def: Aversive Power

- *x* can prevent *p*
- for a given agent *y* and a given goal *q* of *y*, *p* implies not *q*

Note that the previous definitions would require a rich semantic to be really different. In simple logics, since *p* can be only true or false, there is no difference between being preventing *p* and bringing about non-*p*. More complex definitions would require, for example, the usage of temporal modalities. In this case, "prevent *p*" can be read as "always not-*p*", while "bring about *p*" would correspond to "*p* and AGT *x p*?", i.e., *p* is made true by the action of *x*.

3.15 Common intuition would require that if *p* is a social harm, only if *x* has aversive power on *p*, then *x* is responsible over *p*. Indeed, agents may cause events that they are not responsible for (this is the case with children, for example). On the other hand, agents may be responsible for events which they did not directly cause nor decide upon (this is the case of parents with regard to their children's doings).

Power and Responsibility

3.16 How to ascertain *x*'s responsibility? By executing a number of tests on possible external (exogenous) and internal (endogenous) sources of *x*'s aversive power.

3.17

Exogenous power. Given *s*,

- *Is there any agent x which is empowered* (cf. Jones and Sergot 1996) to control or avoid s? Empowerment has important effects on responsibility. Agents are responsible for the social harms which they are empowered to prevent. For example, a guard in an art museum has a special power on visitors, to prevent them from getting close and touching the paintings. If a visitor spoils one of the paintings, the guard will be held more responsible than any other witness, just because she had more power than anybody else in the same setting to avoid *s*. However, empowerment has to do with task-based responsibility, and will be examined later.
- *Are there non-institutional social factors implementing x's capacity to prevent s*? These essentially consist of *x*'s status or position in the social hierarchy, and of the reputation it enjoys within the group. In part, these factors are based upon *x*'s internal power, but this is not always the case: social status might be inherited, and reputation may be impaired by contingent factors. For example, a newcomer's capacity or abilities may have less deterrent effect than those of known members of the group, other things being equal.
- *Is there any agent which has control over resources involved in* s? Suppose someone injures itself or others with a gun. Who does the gun belong to? The gun owner will be held responsible more than everybody else, other things being equal, for *s*. Resources may have a conditioned negative power: if they are used, they may enable users to cause harm. Consequently, those having control on potentially harmful resources have an aversive power at least proportional to the negative power of the resources they control, since they can prevent them to be used improperly. However, this type of power is a source of responsibility only if the agent is also attributed the mental power to predict events.
- *Are there situational factors implementing x's capacity to predict s* (situated cognition)? Agents are responsible for the effects that their situation allows them to predict. In his work on monitoring execution in teamwork, Kaminka and Tambe (1998) gives interesting examples of how agents may be favoured in predicting possible mistakes of their partners in teamwork, by their individual perspective, which we call here situated cognition. The more an agent's predictive capacity is favoured by its "point of view", the higher its aversive power, and consequently the higher its responsibility if the predictable mistake occurs.

3.18

Endogenous power. Given *s*,

- *Is there any agent x with the physical capacity to prevent s?* This amounts to a subset of *x*'s physical properties (such as, strength, health, etc.),
- *Is there any agent x with the mental capacity to predict and prevent s*? The latter mainly consists of *deliberative capacity*, which includes the capacity of predicting events, reasoning about and deciding about them (the epistemic power that Aristotle attributes to responsible agents). One of the ways to ascertain agents' deliberative capacity is their level of maturation. Given their lower level of maturational capacity, youngsters are deliberative agents to a lesser degree than adults. Consequently they are less responsible

than grown-ups, or non-responsible at all (usually, their responsibilities are held by their parents, see later on in the paper).

3.19

Power Limits and Extenuating Factors. Once an agent (or a set of agents) with aversive power with respect to *s* is created, its (degree of) responsibility must be ascertained. Given *s*, and one (or more) responsible agent *x*,

- Are there situational constraints limiting *x*'s power to predict/prevent *s*? These may be either pragmatic or epistemic: *x*'s direct intervention to prevent the occurrence of *s* may be materially obstructed under given circumstances (pragmatic constraints). Analogously, situated cognition may represent an obstacle for predicting *s* (epistemic constraints).
- *Has s been caused directly by a deliberative agent y?* Usually, another deliberative agent's intervention prevents *x* from predicting *s*. Suppose John lends his car to Mary and she then has a car accident. If Mary is a deliberative agent, John is not responsible for the consequent injuries.
- *Had x the power to predict that y will have caused s? x* is not responsible for *y*'s mistakes or wrongdoings unless *x* is in the condition to foresee that *y* will cause *s*. In the previous example, John will be held responsible only if he could have predicted s; for example, if he knew that Mary has recently had a nervous breakdown, or if his car is an old one and needs to be driven safely. He should not have let Mary drive the car.
- *Has an agent y exercised coercive power on x?* Agents are less responsible for s when their power to prevent it is limited by the power that others have on them. A bank casher, under the burglar's gun-shot, is not responsible for having told him the combination of the bank's safe.
- *Is a norm prescribing s?* Agents are not responsible for effects *prescribed* by norms. Suppose Mary happens to witness a murder. After her deposition, the murderer is sentenced to death. Mary is certainly not responsible for this epilogue. However, deliberative agents are responsible for the effects that violate other norms, possibly superior to those that they have respected. This implies the capacity to apply a preference order, or to solve norm conflict. Suppose a surgeon must remove the leg of her patient, otherwise this will most certainly die. Is she responsible for the successive pains of her patient? Not quite. Why? Because the surgeon is actually prescribed to act in the global interests of her patient[3]. A merciful surgeon who does not conform with such a prescription, would certainly be held responsible for the consequences of her decision on the patient's fate (this point will be developed later on in the paper). On the other hand, the Nazi criminal who did not rebel against the commands received *is* held responsible for the victims of those commands, indeed (unless he was under a real threat for his life). Why? His decision did not respect what is considered as a universally desirable preference order (which in this case assumes that the *jus gentium* is to be preferred over military duty).

3.20 To sum up, we will say that:
Def: Social Responsibility (power)
an agent *x* is socially responsible for an objective world state *s* iff

- *s* is an objective social harm, although not necessarily predicted nor wanted by *x* nor by the victim,
- *x* has the internal and external power to prevent s, whether *s* is consequent to *x*'s actions or not.

3.21 When these conditions apply, we will say that *x* is responsible for *s* to a degree that existing extenuating factors will help to determine. In Figure 2 we collect some of the factors involved.

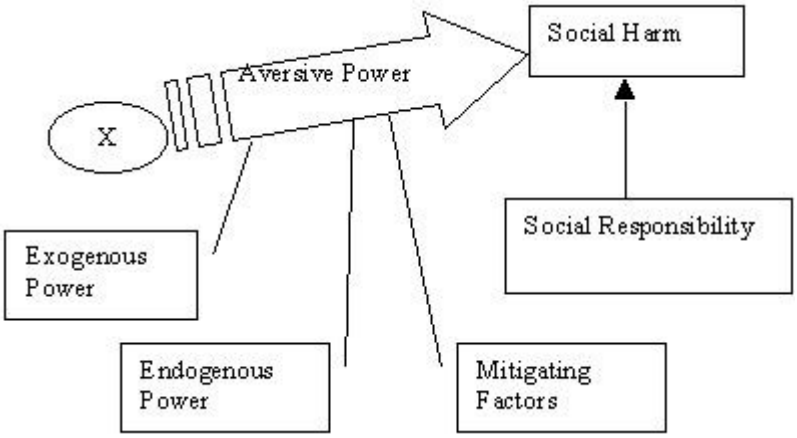

Figure 2. Factors involved in the evaluation of the aversive power of agent x on a social harm that entails social responsibility

3.22 Responsibility is therefore a property emerging from the external evaluation of an agent's capacity to predict and prevent objective social harm. Such a capacity includes the mental power to foresee consequences of actions, the physical power to execute the necessary actions, the social and institutional power on potential harmers and victims, the normative power to solve conflicts according to some socially desirable preference order. An agent is responsible not only for its imprudent behaviour but also for its idiosyncratic preferences, if these have consequences on others' well being. Responsibility is not referred to the agents' (ill-)will, but to their physical power, their mental, moral and social competence, and their institutional empowerment.

Primary Responsibility: An Unremitting Property. Guilty, Accountability and Conflict

3.23 Agents cannot give up their responsibilities at their will. They can only renounce the privileges that entail responsibilities. As we shall see, people avoid taking responsibilities in order to avoid accountability. But there are fundamental responsibilities that can never be dropped voluntarily. This aspect of responsibility, which emphasises its objective nature, strengthens its deterrent force and its social efficacy.

Responsibility and Guilt

3.24 What is the relationship between guilt and primary responsibility? We believe that these two phenomena are strictly related without overlapping: guilt is the responsible cause of harm; responsibility is the power of avoiding it. Of course, an agent who is not responsible for a given harm, cannot be guilty (if she feels she is, she also feels responsible, irrespective of others' evaluations; here, however, we will not analyse the subjective feelings corresponding to responsibility and guilt, but only their objective characteristics). But responsibility does not imply guilt.

Accountability: To Respond for One's Responsibilities

3.25 What are the consequences of responsibility? An agent is accountable for its responsibilities; this leads to *x*'s been asked by the victim and/or the institutional or non-institutional entity supporting the victim to

- give reasons for it; if no justifiable[4] or extenuating reasons are given
- *x* is asked to bear the consequences of social harm in a measure which usually corresponds to its degree of responsibility, and therefore to (contribute to) repair it or provide compensation. Consider the following two situations: (a) *x* has witnessed a rape but did not prevent it, although he had some power to do so; (b) *x* lends her car to a

friend who got no driving license, and who then kills a pedestrian. In (a), *x* is less responsible than in (b). In (a), *x* won't bear the consequences of rape, although he might be asked to respond for his passive behaviour, for not having called the police, for missed succour, etc., while in (b), *x* will be sentenced to bear the costs of repair. Responsibility allows to enlarge the space of accountability beyond causality. Furthermore, it allows to quantify accountability. Thirdly, it allows for the costs of repair to be distributed and be shared within the social environment.

- *x* may bear some punishment (loss of reputation, etc.), and/or
- be revoked privileges entailing responsibilities, and therefore the capacity to damage others (e.g., driving license, bank account, credits, civil rights, liberty, etc.). In particular, the agent may be revoked a general and even institutional trust.

Def: Accountability

an agent *x* is accountable for a social harm *s* when

- x is (socially) responsible for s
- *x*, rather than or together with others, *responds for s*:
  - give reasons for s,
  - bear the consequences (repair), and/or
  - possibly bear some punishment.

3.26 Indeed, accountability leads to an obligation impinging on *x*: to the extent that *x* is found responsible for *s*, it ought to respond for *s*. How to determine the costs that *x* ought to sustain? These are proportional to

- *x*'s responsibility (the extent to which *x* is responsible for s) and
- the extent to which *x* shares its responsibility with others: the smaller the set of agents which are accountable for s, and the greater the costs of repair which each of them is asked to sustain (see below).

Conflict of Responsibility

3.27 Consider the case of a person who perceives that her best friend's husband has an affair with another woman. Should she tell her? Should she "take" such a responsibility even though she is not required to do so? If she does not, she may later be found primarily responsible for the social harm that her friend has received in the meantime ("you let me cut such a bad figure in front of everybody's eyes without telling me!"). On the other hand, if she speaks to her friend, she takes an equally or even more severe responsibility, which may lead to the breakdown of a long-lasting, even happy marriage with all its predictable consequences. How the unfortunate witness will solve her problem is not our business; the point is that people may be expected, and sometimes even prescribed, to take a non-requested responsibility. How is that possible, why do we speak of responsibility in such cases? In order to avoid a given harming effect (loss of reputation etc.), *x* may become responsible for a more serious one (marriage breakdown). Interestingly, *x*'s responsibility depends on the socially desirable preference order. If *x* decides to keep silent and one's loss of reputation is preferred to marriage breakdown by the social group *G*, *x* won't be accountable for her friend's loss of reputation; whereas she will, if *G* values reputation more than a happy marriage.

Task-based Responsibility

3.28 In the following, we will analyse a second notion of responsibility, usually referred to with the expression "to accept a given responsibility". Often this implies a *specific commitment* to a given task (or set of tasks), and sometimes it is a side-effect of a *general commitment* to a given role. A physician is responsible for the health of her patients. But what does this mean more precisely? Does task-based responsibility merely overlap with commitment? How is it related to primary responsibility, if any such link exists? As a preliminary definition, we say that task-based responsibility is a relation among:

- an agent *x*, the responsibility holder,

- a given task *t*
- the entity *z* entrusting *x* (see Castelfranchi and Falcone 1998) with *t* for a goal *gz* in a plan *Pz*.
- The harm *sz* which are caused to *z* from *t* not being executed (in all or in part) or from its being executed wrongly.

3.29 To be noted, *gz* may be a benevolent goal of *z* with regard to *y*. In such a case, the direct beneficiary of *t* is *y*, and the harm may refer primarily to *y,* but *x* is responsible for task execution *before z*. Furthermore, *t* may consist of a series of actions aimed to a maintenance goal: a person responsible for controlling a given equipment must check the integrity and functionality of this equipment.

3.30

Task-based Responsibility and Obligation. Task-based responsibility is based upon obligations: agents are strongly responsible for the effects of an unfulfilled task. Two specific sources of obligations deserve attention.

- *Role adoption*: Often, task-based responsibility is a consequence of role adoption: an agent who accepts to play a given role takes a responsibility with regard to the accomplishment of that role, i.e. with the tasks associated to it. Task-based responsibility does not imply a specific commitment to the task *t*, since this may be assigned to *x* from its role. Parents are entrusted with the task to take their children to maturation.
- *Commitment*: in such a case, one cannot be responsible if one's task responsibility is not agreed upon both *x* by *z*. Suppose that before going out Mum asks who, whether John or Mary, will take out the dog today: both children separately answer they will do that, but Mom decides to entrust John, because Mary is much better at answering the telephone politely. In fact, John forgets about the dog, but is Mary responsible before her mother's eyes if the dog urinated on her new Chinese carpet? Probably not: Mary cannot be held responsible for a task she was not entrusted with. Even more interestingly, Mary's primary responsibility for *s* (spoilt carpet) is lowered by the extenuating factor that the task was assigned to someone else.

3.31

Task-based Responsibility and "Counting-Upon". Primary responsibility deals with a loss of power of the victim. What about task-based responsibility? Apparently, in this case agents are entrusted with tasks designed for achievement goals, rather than maintenance goals. However, task-based responsibility renders agents accountable for the harm which entrusting agents derive from wrong investments, from delegating tasks which will not be fulfilled. Task-based responsibility leads *z* to form the belief (expectation) that *x* will take care of *t*. Consequently, *z* discharges itself from (any concern relative to) *t*. Indeed, *z* will "count upon" *x* for *t*. This is something more than relying upon *x*. In *reliance* (see Castelfranchi and Falcone 1998),

- *z* believes that *x* can do *t*,
- wants *x* to accomplish it,
- believes that *x* believes that *z* relies upon *x*, and finally
- will not ask anyone else to accomplish *t*

.

3.32 In *counting-upon*, some further mental states of *z* should be added, namely *z* believes (expects) that

- *t* will be fulfilled and
- *z* will not sustain further costs of task execution (beyond that already sustained for "hiring" *x*).

In other words, *x* responsibility about *t* leads *z* to take the fulfilment of *t* for granted, to form expectations about it. A disconfirmation of this expectation would represent a multiple loss for

z: (a) *delegation costs*: the costs implied by searching and hiring *x* will have no compensation; (b) *opportunity costs*: the costs implied by giving up other solutions, which might be presently unavailable; (c) *investment costs*: the costs implied by the compromise of *gz* plus the compromise of further goals which have the result of *tx* as a condition.

3.33 Based upon these concepts, we establish the following
Def: Task-Based Responsibility

- *x* is responsible before *z* for the fulfilment of *t*, when
- due to the role which *x* plays in a multi-agent plan *Pz* for a goal *gz* of *z* (which may be benevolent with respect to a direct beneficiary *y*, and in such a case *x* is assigned by *z* the task to benefit *y*), or to its specific commitment to *t* before *z*, there is an obligation on *x* to accomplish task *t*
- *z* counts upon *x* accomplishing *t*:
  - *z* believes that *x* has obligation to accomplish *t*,
  - *z* wants that *x* accomplishes the obligation and therefore the task *t*,
  - *z* believes *x* will do so491 ,
  - *z* believes that *x* believes that *z* counts upon *x* for *t*, and
  - *z* believes *t* will be accomplished without further costs for *z*.

The ingredients for task-based responsibility are shown graphically in Fig. 3

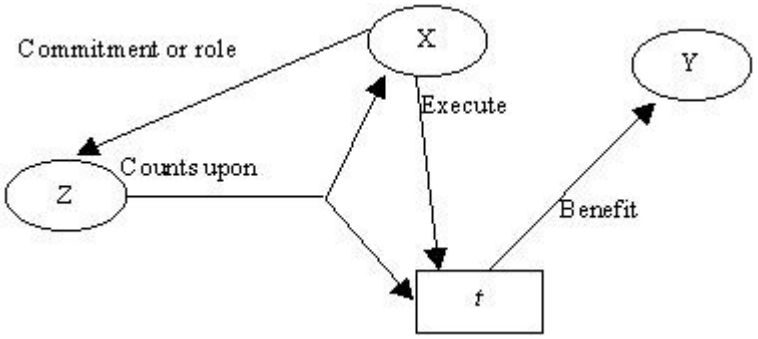

Figure 3. A graphical representation of task-based responsibility

The Confines of Task-Based Responsibility

3.34 Suppose Mary commits to take care of a urgent fax that her colleague John is unable to send out on his own. She then accepts a task-based responsibility with regard to the fax. How far does one's task-based responsibility extend? In some sense, this depends on the specific commitment (Hart and Honoré 1985): if Mary accepts to send out John's fax, Mary is held responsible for a thorough accomplishment of the plan "send fax". This usually includes that the sender *checks* whether the whole fax goes through (an OK receipt is printed from the fax machine with the number of pages actually transmitted), and recursively applies the send-out procedure until this print is available. Therefore the task "send fax" actually implies that two subprocedures be executed: one for sending the fax (put the sheet on the machine, checks that the machine is ready, dial the number, etc.), and another for checking that the fax gets through. If Mary gets away without checking whether the fax got through, she did not accomplish her task, and is accountable for the consequences of non-fulfillment.

3.35 A more extensive responsibility that goes beyond the shared script for sending faxes would imply an explicit negotiation and agreement. For example, John should ask Mary not only to check whether the fax goes through integrally, but also to give a call to the recipient and ask him whether he got it in a readable form. Only if Mary commits to this much, she can be held responsible for any omission in this more complex task.

Interplay (or Trade-Off?) Between Primary and Task-Based Responsibility

3.36 It is interesting to observe that the execution of a task usually exposes *x* to face novel primary responsibilities. If, while accomplishing her promise to John, Mary damages the fax machine, under ordinary conditions Mary is the person to be blamed rather than John (unless, in the firm they both work in, John is "responsible for" the fax machine).

3.37 Moreover, sometimes the two types of responsibility may be incompatible, and require conflict management and solution. This is somehow obvious, since there may be incompatibilities among the prevention of different harms. To prevent *si* (primary responsibility) may be incompatible with preventing *sj* (task-based responsibility). Suppose that John accepts Mum's request to take out the dog; but when he is just about to get out, John notices some bad guys standing in front of his house. Will he leave the house and his young sister Mary alone in such a dangerous situation? Certainly not. What is more interesting is whether in such cases, an agent is accountable for the consequences *sj* that will occur if *x* chooses to prevent *si*. Is John accountable for the dog's later wrongdoing? Probably not, but he would certainly be for the effects of burglary if he chooses to walk away. The solution of the puzzle resides in the entity of the harm: agents are accountable for the *worse* consequences, i.e. for harms that are worse than those they prevent. Agents may, or are even expected to, break a task-based responsibility in order to fulfil a primary one provided the consequence of breaking the task-based expectation is (prescribed to be) preferred over the consequence of breaking the latter. People's responsibilities are continuously assessed and evaluated in terms of how socially acceptable the decisions they take -- while accomplishing their tasks - are (Lacey 2001). The Nazi officer who kills thousand people to meet his commitment to the German army is certainly found responsible for the victims' lives. He is actually expected and prescribed to break one's commitment in view of a socially established higher-level responsibility. Therefore, one can be held primarily responsible for accepting a task-based responsibility[5]!

3.38 Secondly, and more crucially, one's acceptance of a task-based responsibility may lead one to "take" new task-based responsibilities which *x* will have to respond for later. Suppose that the fax machine is found out of order: neither John nor Mary were informed (or perhaps John was, and took advantage of Mary by leaving her with that problem...). John, in the meantime, is up and away. What should Mary do? Should she go out and look for another fax machine in the neighbourhood? Or should she "take" a further responsibility, e.g. to put the fax in an envelope and send it by s-mail? What if the fax gets lost by the mail service, or John is not allowed to resort to s-mail? What kind of consequences is Mary accountable for? On the other hand, is she allowed to put John's fax aside and forget about it? What types of consequences is she facing in this case?

3.39 In Fig. 4, we try to arrange graphically all the concepts involved in our two definitions of responsibility in order to show similarities and differences between the two.

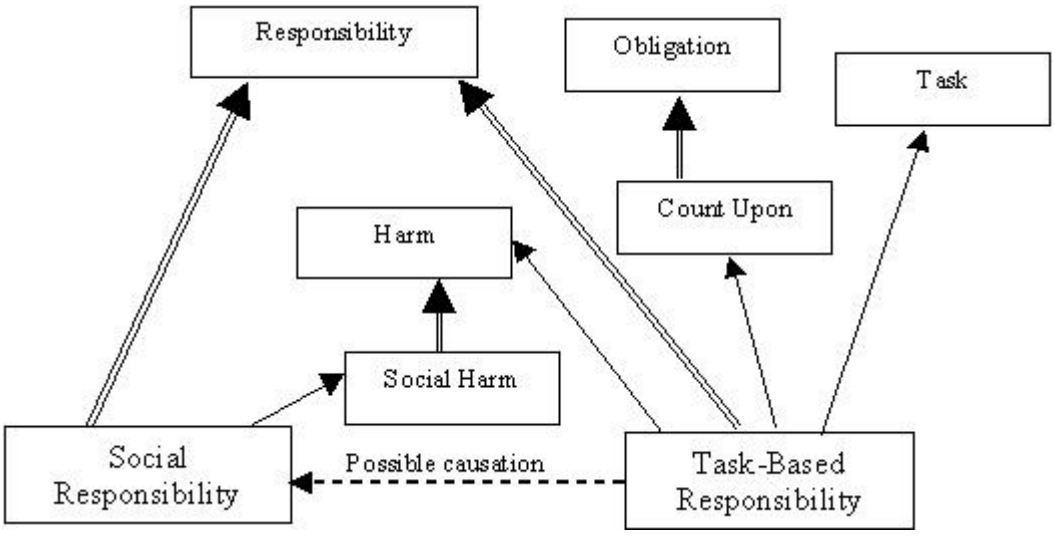

Figure 4. The two faces of Responsibility. The dependence (single line) of Social and Task-Based Responsibility is shown graphically, together with generalisations (double line) of cognitive constructs employed

Vicarious responsibility

3.40 As we have seen in primary responsibility, agents might be accountable for social harms caused by others. More generally, agents may have primary responsibility *in place of* some other agents that are entrusted to them and are not attributed aversive power. In other words, agent *x* is responsible for and in place of an agent *y* when *y* is not attributed the power to avoid harming itself or others, while *x* is attributed such aversive power and *y* is entrusted to *x*. A parent is responsible for her children causing harm to others or themselves. Conceptually, the minimal notion of being entrusting to someone resides in vicarious responsibility.
Def: Vicarious Responsibility
x is responsible for a given agent *y* when
- *y* is not attributed the power to prevent harming itself or others, while *x* is
- and *x* is accountable for harming effects on or caused by *y*.

In vicarious responsibility, *x* respond on behalf of, or in place of, *y*. Stated in Jones and Sergot's (1996) terms, its actions *count-as y*'s.

3.41 Often, the concrete phenomenon of vicarious responsibility implies a "tutorial" role of *x* with regard to *y*, namely with the goal of watching over *y*'s interests, influencing *y*'s goals and actions in *y*'s interest. But this is not necessarily the case. An artificial agent might be entrusted to its designer (or user) in this minimal sense: only the latter is accountable for the effects of the artificial agent on others.

Task-Based Accountability

3.42 To accept responsibility implies that agents may be called to respond for the world state *s* which is consequent to a partial, null or wrong execution of their tasks.

3.43 Which world states is *x* responsible about and before whom? Only before *z*, the entrusting entity? Two fundamental types of world states should be considered:
- Harming effects of a null, wrong or partial execution of *t*: this is a responsibility which *x* holds *before z*, the entrusting entity which counts upon *x*, and which would be damaged by *x*'s insufficient or inadequate execution.
- Harming effects contingent to or caused by task-execution: these are included in *x*'s primary responsibilities, which *x* holds *before the potential victims* and, more generally, *before the social group*.

3.44

x must explain why a certain effect occurred, whether it has accomplished the task it was charged with, why it decided to accomplish it in a given way, why it did not accomplish it at all, and possibly repair some of the consequences of task non-fulfilment, and finally be even removed from its responsibilities.

Def: Task-based Accountability

x is accountable for *s* in *t* when

- *z* counts-upon *x* for *t*:
  - *z* believes *x* can do *t*;
  - *z* believes that there is an obligation on *x* to accomplish *t*;
  - *z* believes that *x* will want to do *t*;
  - *z* wants *x* to accomplish *t*;
  - *z* believes that *x* believes that (d);
  - *z* believes that *t* will be accomplished; and as a consequence,
  - *t* non-fulfilment is a sz (a harm at z's expense).
- *z* is authorised to expect/exact that *x* responds for sz caused by *t* non-fulfilment.

3.45 Given the considerations made in the preceding subsection, what are the factors contributing to estimate *x*'s task-based responsibility?

- *Task specification*: if *x* did not commit to check the content of the fax, it will not be held responsible for the loss of reputation which John obtains from an incorrect text of the fax. However, the type of commitment may act also as a strengthening factor. To accept an open delegation (have the fax at destination in due time; cf. Castelfranchi and Falcone 1998) may lead *x* to respond not only for the effects of task *t'* (send a fax), but also for those of alternative plans (send the text by s-mail).
- *z's task-based responsibility with respect to t*: suppose that John is responsible for the fax machine management and use. If Mary breaks the fax machine, *z* will be found as much, or even more, responsible than her. This follows from the present analysis of responsibility as an unremitting property. If *z* has a responsibility over a given *t*, it may delegate it or entrust *x* with it, but it is *z* which will be called to respond for *s* associated to *t*, even if *s* is caused by another agent.
- *The organisational structure and hierarchy*: *x* is not responsible for the effects of *t*'s execution on the whole organisational plan, although it remains primarily responsible for the harming consequences of *t* with regard to other, more general interests (norms and public interests).
- *z's power on x*: as said before, *x* cannot be held responsible for some task that it is obliged to accomplish. However, one is responsible for effects of that task in a primary sense, if these are socially valued as worse than the effects of task non-fulfilment.

Multi-Agent Responsibility

4.1 Responsibility can be either individual or multi-agent. In the latter case, it can be either shared with others or collective.

Shared Responsibility

4.2 Passengers witnessing a murder or a rape without intervening share a responsibility with regard to it. All of them are accountable for *s* to a degree that depends on the entity of the social harm, on the type and amount of their power, and on the number of agents which share this power. In these conditions, responsibility, and therefore accountability, are equally shared among the agents. One major social consequence of a shared responsibility is that each agent is accountable only for a share of the social harm: the more the number of agents that share it, the lesser the individual contribution to the social harm, and the lesser the individual accountability[6]. This is in fact one of the reasons why the higher the number of attendants,

the lower the probability that any of them will provide help in emergency, an effect which has been observed by social psychologists since long (Latané and Darley 1972). But what does shared responsibility actually mean?

4.3 Consider the case of agents acquiring dubious merchandise (for example, too cheap): these might share a responsibility with regard to the exploitation of children's work. None of the customers is sufficient to prevent this undesirable social consequence, however each of them might act independent of others: the exertion of aversive power does not imply interdependence, while the *efficacy* of power implies that this power is actually exercised by all members of the set. The classical free-riding problem is an example of this type of responsibility, which is necessarily shared by the agents although each of them might contribute its share of the public good independent of agreement and cooperation with others. In these situations, complementarity is relative to the effects of action, rather than to its execution. An analogous situation is found in traffic jams: everybody contributes, but everybody could do something independent of others. Of course, the result is significant and perceptible only if everybody contributes. Def: Shared Responsibility Given a social harm *s*, and a set of agents X, the members of X are attributed a shared responsibility for s, when

- each *x* in X is attributed the power to reduce *s*
- each can exercise its aversive power independent of others.
- the avoidance of *s* is proportional (although non-linearly) to the amount of contributors

.

Collective Responsibility

4.4 There are circumstances in which a set of agents, rather than sharing responsibility, bear a collective responsibility. What does this mean? Consider the (Jennings and Mamdani 1992; Norman and Reed 2002) example of two children which Mum asks to clean their rooms: if either will be found still messy on her return, both children will be held responsible *no matter whom* the messy room belongs to. Or else, consider the case that John asks his colleagues to send a fax: both Mary and Ron answer that they will take care of that but they don't know yet *which one* will actually send the fax[7]. John holds both agents responsible, although only one is needed to send the fax. Agents are collectively responsible for the task execution because they cannot act independent of each other. It is up to them to decide (see again, Norman and Reed 2002) which does what, but they must take into account the other. If Mary decides that she would rather stay home the day the fax is due, she ought to call Ron, and negotiate with him upon who is the one which sends the fax: if Ron accepts, than Mary stays home and Ron sends out the fax.

4.5 Therefore, it is not necessary that all members of *X* actually execute the task *t*. What is required is that all of them want that *t* to be executed by *X* or some of its members.

4.6 This goal is a typically common goal (see Conte et al. 1991) with regard to which *X*'s members are intrinsically interdependent, since it requires a multi-agent plan:

- execution: this sub-plan is decomposed in several actions: find the agents which can do the task (and if alternatives exist, choose according to some criterion), have the task accepted by executors and task-allocation agreed upon by other members;
- check execution: members of *X* will not achieve their goal until they know that
  - *t* is completed: if executors did not accomplish it, the collective will have to substitute them (and the previous sub-plan be applied until *t* is executed).
  - *t* is effectively executed; that is, errors are avoided[8].

4.7 Of course, these sub-plans may be executed on the grounds of some pre-existing organisational structure of interdependent roles (executor and controller), which *X*'s members instantiate.

Def: Collective Responsibility

X is collectively responsible for task *t* in front of *z* when

- *z* counts upon X for *t*:
  - *z* believes that at least a non-empty subset of X CANDO *t*;
  - *z* wants that all *x* in X want
  - that *t* be done;
  - *z* believes that all *x* in X want that *t* be done;
  - *z* believes that all *x* in X have an obligation to want that *t* be done;
  - *z* believes that *t* will be done;
  - *t* non-fulfilment is a loss for *z*.
- all *x* in X have a common goal that *t* be effectively executed by (a subset of) X's members.

Collective Accountability

4.8 Given a set of agents *X* collectively responsible for *t*, and given a social harm *s* consequent to errors or omissions in *t*'s execution, which agent in *X* will be accountable for *s*? Whom shall the victim (the entrusting entity *z*) address itself to in order to obtain repair? Which agent in *X* will bear the effective consequences?

4.9 The answer is a direct consequence of the previous analysis of collective responsibility. If responsibility is collective, accountability will also be collective. In particular, the whole set of agents *X* will face the consequences: it will be *X* itself which will decide whether and how to redistribute the consequences of errors and omissions of its members. While individual and shared responsibility indicate the individual locus (or loci) of accountability, collective responsibility indicates a collective locus of accountability, within which effective costs may be distributed according to criteria and degrees which are internal to the collective. This implies that the collective can be addressed to provide repair, but how and which concrete agents will actually contribute to this and to what extent, is determined by the collective itself.

4.10 In this sense, the collective action's opacity to external control is much higher than individual or shared actions. Collective action may diminish or disguise individual responsibility. From the outside, it is impossible to say which agent is effectively responsible and accountable for which effect and to what extent. The collective action represents a double filter: it filters both task-allocation and accountability.

Def: Collective Accountability

A set of agents X is collectively accountable for a given social harm *s* in the execution of a task *t* when

- X is collectively responsible for *t*
- *z* counts-upon X for *t*
- *z* is authorised to expect/exact that
  - some members of X respond for *s*,
  - bears the costs of repair and/or
  - some punishment.
- X's members decide which ones among its members will effectively bear the above costs, and whether and how these should be distributed

.

Domains of Application

5.1 Two distinct advantages of the present analysis can be envisaged with regard to ICT domains of application.

Responsibility and e-Governance

5.2 In general responsibility and other objective social notions (for example, reputation) facilitate what we call *e*-governance, i.e. the extension of indirect forms of regulation, not necessarily implying the issuing of norms, to individual and corporate users of information and communication technologies.

5.3 A second more specific advantage concerns the governance of agent-mediated interaction, transaction and negotiation. Reputation-based systems already operate in this sense, not only in the domain of *e*-commerce, but also in teamwork for different applications. Responsibility for context-relevant social harm may be easily attributed to autonomous, automated systems in multi-agent contexts (whether these interact with humans or other automated systems) thereby allowing artificial *loci* of accountability to be identified. An interesting connection between accountable agents and reputation-based systems can easily be perceived. Individual and artificial agents found (co)responsible and (co)accountable for given social harm, are at risk of an endangered reputation, what may have a positive impact on the social knowledge of potential partners. In particular, to identify the bases of the responsibility power, its limits and the extenuating factors might help produce protocols for responsibility attribution with regard to defined and context-relevant harms.

5.4 In turn, this analysis and the instruments that might be built up as a consequence of it seem also to bear a positive, improving impact on the successive performance of responsible agents, whether individual or collective.

Responsibility and Organisational Structure

5.5 We will conclude this paper by mentioning the potential utility of the present analysis for exploring organisational structures. Both the notions of responsibility as an unremitting property and that of counting-upon may be useful in the process of designing an organisational structure, and conversely in understanding how a given organisation is structured. To give but one obvious example, in horizontal organisations, where task-executors occupy equal hierarchical positions, the number of *loci* of responsibility and accountability are a direct function of the number of task-executors. In hierarchical organisations, with nested structures and task sub-delegation, the loci of responsibility are not necessarily transparent.

5.6 Both structures have advantages and disadvantages. In horizontal organisations, responsibility maybe distributed and diluted among task-executors. But, once identified, the responsible agents will directly negotiate with *z* about whether and to what extent they will repair/respond for a given *s*. With sub-delegation, things are different. Whether it is nested within the organisation (hierarchical structure) or external to it (in open organisations with outsourcing), responsibility is not diluted, but the process for finding and negotiating with the agents accounting for a given *s* is not entirely transparent. This has to do with the phenomenon of *hidden* responsibility, which is more likely to depend upon nested collective responsibility. The preference of diluted over hidden responsibility or vice versa should be explored (possibly by means of experimental computer simulation) with reference to different types of tasks and other organisational variables.

Concluding Remarks and Future Work

6.1 In this paper, we endeavoured to provide a model of responsibility based upon *aversive power*, rather than decision and action. This allowed us to answer a number of questions raised at the beginning. In particular it allows us to

- clarify the interconnections between responsibility, causation, obligation and guilt
- clarify the interconnections between responsibility and accountability
- provide a preliminary notion of *counting-upon*, as a fundamental aspect of role-based responsibility,

- provide a general elementary definition of responsibility, in its interconnection with a more specific, task-based one
- do justice to two distinct intuitions.
    - Responsibility has a historical, dynamic content (determined by the variability of the notion of *social harm* and of the moral preference order). Historical variability is not incompatible with the emergence of objective responsibility. Possibly, any society at any step of its evolution may require and give rise to such a property, whatever its culture-specific notion of social harm and the preference order among possibly conflicting harms.
    - The range of co-responsibility is larger than teamwork: agents can be found co-responsible for world states about which they took no decision.

Examples of the potential of this analysis for the study of organisational structures and for the governance of e-societies have been discussed. In future extensions of this work, we plan to formalise the basic concepts exposed here following some of the most common approaches for logic-based agency, with the possible worlds semantic that was implied in our pre-formal analysis. However, in order to account for quantification of responsibility, we will also need to reconcile the logic-based approach with some kind of numeric measurement. In addition, we plan to examine the interaction of social responsibility with several other cognitive constructs which we consider fundamental in social analysis, like trust and reputation.

Acknowledgements

This work has been partly supported by the FIRMA (EVK1-CT1999-00016) project. We would also like to thank the anonymous reviewers for their helpful comments.

Notes

[1] See the Centre for Computing and Social Responsibility (CC-SR), http://www.ccsr.cse.dmu.ac.uk/

[2] But there are others. A socially relevant harm is one that the social group, which both the victim and the harmer belong to, is interested in avoiding, since it endangers some global goal or interest of the group. This point requires some clarification. A group may have an interest that (a) its members are in the proper condition to exercise their social role/function and therefore that (b) this condition be maintained or (c) restored at the lowest cost for the whole group; consequently, (d) the group is interested in avoiding that the costs of restoring their members' capacity be sustained by the whole group, and that it is (e) distributed over a subset of the group. Social groups are interested in charging a given number of agents with the costs of repairing social harms. Hence, they assess which agents have effective responsibilities for given social harms. At the same time, they are also interested in avoiding injuries and in reducing the social (whether global or distributed) costs of repair. Consequently, they are interested in discouraging socially impairing behaviours. Hence, they claim that responsible agents exert their power to avoid potential harms.

[3] Indeed, she would be held responsible for a decision which did not take into account the global interest of her patient.

[4] In Gardner (2003), responsibility is indeed defined as the ability to respond, i.e., to give justification of given choices; however, this view is not satisfactory, fist because makes

responsibility collapse on accountability; secondly, because it still is a decision-based notion.

[5] However, harm comparison is not always an easy one. Suppose a physician refuses to practise infibulation in a young patient of Islamic religion in virtue of his commitment to defend the physical integrity and the health of his patients. Can he be held responsible for the severe injuries that a non-institutional infibulation may cause to the child? The answer is not an easy one.

[6] Some (e.g., Mellema 1988) consider this dilutionist view of shared responsibility as somewhat unsatisfactory. In contrast with it, an anti-dilutionist view is proposed. As regards shared responsibility, the author seems to argue that all bear responsibility. As regards collective responsibility (Mellema 2001) it may be the case that some members of the collective do not bear responsibility, while the whole entity does.

[7] This example indicates a possible solution to a classical problem in deontic logic: what is a collective obligation? More precisely, how to predict which agents is the obligation impinging upon? The two identified solutions (cf. Carmo and Pacheco 2000) are complementary: either the obligation actually impinges upon at least one agent in the collective, or on all of them. Both solutions have drawbacks (for a convincing critique, see again Carmo and Pacheco 2000). A possible way out is allowed by distinguishing conceptually an obligatory goal and the consequent obligatory action: a collective obligation is one that all members of the collective ought to want to be realized, although only a subset of agents are sufficient to execute it.

[8] Sometime, X's members may be interdependent in avoiding errors: to see this, consider Kaminka and Tambe's work on monitoring teamwork execution (1998). As is reasonably suggested in that work, avoidance of errors in teamwork execution is facilitated by decentralized control not only because the higher the number of agents who effectuate control the lower the chances of errors, but also and moreover because the team members' viewpoints (and consequently their capacity to predict errors) are different and complementary.

References

ARENDT, H. (1969) 'Collective Responsibility' in James Bernauer (ed.) *Amor Mundi: Explorations in the Faith and Thought of Hannah Arendt*, (Dordrecht: Martinus Nijhoff 1987).

J. CARMO, J.and Pacheco, O. (2000) Deontic and action logics for collective agency and roles, in: *Proc. of the Fifth International Workshop on Deontic Logic in Computer Science (Deon'00)*, R. Demolombe and R. Hilpinen (eds.), ONERA-DGA, 93-124

CARMO, J. and Pacheco, O. (2001) Deontic and Action Logics for Organized Collective Agency Modeled through Institutionalized Agents and Roles. *Journal Fundamenta Informatica*, Vol. 48 (2,3):129-163.

CASTELFRANCHI, C. and R. Falcone (1997) From Task Delegation to Role Delegation. *AI*IA* 1997: 278-289.

COHEN, P.R. And Levesque, H.J. (1990), Intention is Choice with Commitment, *Artificial Intelligence* 42, 231-261.

CONTE R., Miceli M. and Castelfranchi C. (1991) Limits and Levels of Cooperation:

Disentangling various types of Prosocial Interaction, in Demazeau Y. and Muller J. (eds.), *Decentralized A. I. 2*, Elsevier, 147-157

EHRENBERG, K. (1999). Social Structure and Responsibility*, Loyola Poverty Law Journal*, 1-26.

FEINBERG, J., (1970), *Doing and Deserving: Essays in the Theory of Responsibility.* Princeton: Princeton University Press

FLEURBAEY, M. (1995) Equality and responsibility, *European Economic Review*, vol. 39, 3-4:683-689

FRENCH, J.R.P. and B. Raven (1959) 'Bases of Social Power' *Studies in Social Power*. Ed. Dorwin Cartwright. University of Michigan, Ann Arbor.

GARDNER, J. (2003) The Mark of Responsibility. *Oxford Journal of Legal Studies* 23(2), 157-171.

HART, H. L. A. and Honoré, T. (1985) *Causation in the Law*. Second Edition. Oxford Univ. Press.

HEIDEGGER, M. (1977) *Being and Time*; trans. by Stambaugh, Joan; State University of New York Press.

HEIDER F. (1958) *The psychology of interpersonal relations*. New York: Wiley.

HILD, M. and Voorhoeve, A. (2001) Roemer on Equality of Opportunity, Working paper of the California Institute of Technology, Division of the Humanities and Social Sciences, n. 1128, http://www.hss.caltech.edu/SSPapers/wp1128.pdf

JENNINGS, N. R. and Mamdani, E. H. (1992). Using Joint Responsibility to Coordinate Collaborative Problem Solving in Dynamic Environments, Proc of 10th National Conf. on Artificial Intelligence (AAAI-92), San Jose, USA 1992, 269-275.

JENNINGS. N.R. (1992) On being responsible. In Decentralized A.I. 3, *Proc. MAAMAW-91*, 93-102, Amsterdam, The Netherlands 1992. Elsevier Science Publishers, http://citeseer.ist.psu.edu/jennings92being.html

JONES, A.I.J. and Sergot, M.J. (1996) A formal characterisation of institutional power, *Journal of the JGPL*, 4(3): 429-445.

KAMINKA, G., and Tambe, M. (1998) Social comparison for failure detection and recovery. In *Intelligent Agents IV: Agents, Theories, Architectures and Languages (ATAL*), Springer Verlag.

LACEY, N. (2001) *Responsibility and Modernity*, http://www.law.nyu.edu/faculty/workshop/fall2001/lacey.pdf

LENK, Hans (1997) *Einführung in die angewandte Ethik*. Stuttgart Berlin Köln: Kohlhammer.

LANE, M. (2004) Autonomy as a Central Human Right and Its Implications for the Moral Responsibility of Corporations, in T. Campbell and S. Miller (eds.) *Human Rights and the Moral Responsibilities of Corporate and Public Sector Organisations.* Series: Issues in Business Ethics, Vol. 20, Springer

MAMDANI, E.A. and Pitt, J. (2000) Responsible Agent Behavior: A Distributed Computing Perspective, *IEEE Internet Computing*, 4 (Sept./Oct.):27-31.

MELLEMA G. (1988) *Individuals, Groups, and Shared Moral Responsibility*. Bern: Peter Lang

MELLEMA, G.F. (2001), *Collective Responsibility*, Barnes & Noble.

MITCHAM, C. and R. von Schomberg (2000) The Ethic of Engineers: From Occupational Responsibility to Public Co-responsibility. In P. Kroes and A. Meijers (eds.) The empirical turn in the philosophy of technology*, Research in philosophy and technology*, vol. 20, JAI Press, Amsterdam.

MOORE, M. (1998) *Placing Blame*, Oxford: Oxford University Press,.

NORMAN, T. J. and Reed, C. A. (2001) Delegation and responsibility. In C. Castelfranchi and Y. Lesp¥ erance, editors, *Proceedings of the Seventh International Workshop on Agent Theories, Architectures, and Languages*, Lecture Notes in Artificial Intelligence. Springer-Verlag, pp. 136-149.

NORMAN, T.J. and Reed, C. (2002) Group Delegation and Responsibility, C. Castelfranchi and W.L. Johnson (eds.) *Proceedings of the First International Conference on Autonomous Agents and Multi Agent Systems* (AAMAS), Bologna, July 15-19, ACM Press.

OREN, T.I. (2000); Responsibility, Ethics, and Simulation; *Transactions of the Society for Modeling and Simulation International*, Vol. 17, No. 4 pp. 165-170.

SANTOS, F. , A. J. I. Jones, and J. Carmo. (1997) Responsibility for action in organisations: A formal model. In G. Holmstrom-Hintikka and R. Tuomela, editors*, Contemporary action theory: Social action*, volume 2, pp. 333-350. Kluwer.

SARTRE, J.P. (1984*) Existentialism and Human Emotions*, Lyle Stuart.

Scanlon, T. M., (1998) *What We Owe to Each Other.* Cambridge, MA: Harvard University Press

SEEGER, M.W. (2001), Ethics and Communication in Organizational Contexts: Moving From the Fringe to the Center, *American Communication Journal*, Vol. 5(1) http://www.acjournal.org/holdings/vol5/iss1/special/seeger.htm.

SHAVER, K. G., and Drown, D. (1986). On causality, responsibility, and self-blame: A theoretical note. *Journal of Personality and Social Psychology*, 4, 697-702.

SILVER, S. (2002) Collective Responsibility and the Ownership of Actions, *Public Affairs Quarterly* 16 (3) 287-304.

STRAWSON, P.F. (1974) Freedom and Resentment. *Proceedings of the British Academy 48*; Reprinted in *Freedom and Resentment and Other Essays*. Oxford 1974, pp. 1-25. References are to the reprinted version.

ZSOLNAI, L. (forthcoming) *Ethical Decision Making: Responsibility and Choice in Business and Public Policy*, Ashland: Purdue University Press.

Return to Contents of this issue