A logical analysis of responsibility attribution : emotions, individuals and collectives

Emiliano Lorini, Dominique Longin, Eunate Mayor

A logical analysis of responsibility attribution: emotions, individuals and collectives

Emiliano Lorini, Dominique Longin and Eunate Mayor
IRIT, Toulouse University, France

Abstract

The aim of this paper is to provide a logical analysis of the concept of responsibility attribution; that is, how agents ascribe responsibility about the consequences of actions, either to themselves or to other agents. The paper is divided in *two parts*. The *first* part investigates the importance of the concept of responsibility attribution for emotion theory in general and, in particular, for the theory of attribution emotions such as guilt, pride, moral approval and moral disapproval. The *second* part explores the collective dimension of responsibility attribution and attribution emotions, namely the concepts of collective responsibility and collective guilt. The proposed analysis is based on an extension of the logic STIT (the logic of "Seeing To It That") with three different types of knowledge and common knowledge modal operators depending on the time of choice: before one's choice, after one's choice but before knowing the choices of other agents, and after the choices of all agents have become public. Decidability of the satisfiability problem of the logic is studied in the paper.

1 Introduction

The main object of this paper is to provide a framework for the discussion on the complexities and ambiguities that surround the concept of responsibility attribution. A wide range of different, though connected, ideas is covered by this concept.

How and when individuals ascribe responsibility to themselves and to others is a central question in social psychology [47]. Likewise, the concept of responsibility attribution is a core concept in the study of emotion. Particularly, there exists a specific class of emotions, called "attribution emotions" [37, p. 134], which arise when an individual ascribes responsibility to herself or to someone else for a morally deplorable action (*i.e.*, blameworthiness) or for a morally admirable action (*i.e.*, praiseworthiness). Examples of this type of emotions are guilt and reproach. Responsibility attribution is also a relevant concept in the domain of autonomous agents and multi-agent systems (MAS), in particular in the areas of artificial organizations, normative MAS and intelligent virtual agents (*e.g.*, embodied conversational agents, tutoring agents, etc.), where they prove to be useful. For instance, in the case of autonomous agents interacting in the context of an artificial organization, agents should be endowed with the capability to reason about their own responsibility and that of others. This kind of capability allows agents to identify those actions that might be blameworthy, because they do not conform to the organization's norms, and therefore refrain from performing them. More-

over, an intelligent virtual agent interacting with a human can be designed to recognize this human's emotions such as guilt or pride and to act consequently. This specific capacity can be achieved by endowing the agent with the more general capability to reason about the human's responsibility and the human's beliefs about her own and others' responsibility.

Our analysis of responsibility attribution is based on an extension of the logic STIT (the logic of "Seeing To It That") with three different types of knowledge modal operators. Each type of knowledge is defined with respect to the time of the agent's choice: before one's choice (*ex ante* knowledge), after one's choice but before knowing the choices of others (*interim* knowledge), and after the choices of all agents have been made public (*ex post* knowledge). The syntax and the semantics of this logic are presented in Section 2. In Section 3 we show that three different kinds of knowledge operators are necessary for a fine-grained analysis of the different aspects of the concept of responsibility (*e.g.*, the distinction between active and passive responsibility, and the distinction between causal and agentive or moral responsibility). On the basis of the concept of responsibility defined in Section 3, in Section 4 we provide a systematic analysis of attribution emotions; that is, emotions such as guilt and reproach that are based on the attribution of responsibility to oneself or to others. The second part of the paper explores the collective dimension of responsibility attribution and attribution emotions, namely the concepts of collective responsibility and collective guilt. To this aim, in Section 5 we present an extension of the logic of Section 2 with three different types of common knowledge modal operators (corresponding to the three previous individual knowledge modal operators). Decidability of the satisfiability problem for our logic is studied in Section 6. Finally, related literature and future work are discussed in Sections 7 and 8, respectively.

2 Epistemic STIT logic with *ex ante*, *interim* and *ex post* knowledge

STIT logic (the logic of *Seeing to it That*) [6] is one of the most prominent formal accounts of agency. It is the logic of sentences of the form "agent i sees to it that φ is true". In [27] Horty extends Belnap et al.'s STIT logic with operators of group agency in order to express sentences of the form "group J sees to it that φ is true". Following [35], throughout the article we use the terms 'individual STIT logic' and 'group STIT logic' to designate respectively Belnap et al.'s STIT logic (in which only the actions of agents are described) and Horty's variant of STIT logic (in which both actions of agents and joint actions of groups are represented)[1].

In this section we present an extension of atemporal group STIT that allows us to represent the different types of agents' knowledge defined with respect to the time of choice (*i.e.* the agent's knowledge before one's choice, after one's choice, or after the choices of all other agents). We call this logic E-GSTIT (*Epistemic Group* STIT).

STIT has a non-standard branching-time semantics based on the concepts of *moment* and *history*. As shown by [5, 35, 25], however, both individual STIT logic without time axiomatized in [6, Chap. 17], and 'atemporal group STIT' (*i.e.*, group STIT without time) can be 'simulated' in standard Kripke semantics. A similar idea is proposed by [30], who introduce the concept of 'consequential models'. These consequential models are equivalent to the Kripke atemporal group STIT models used by [25, 35], in which the authors abstract away from the

[1] Other authors [8, 45] use the term 'multi-agent STIT' instead of 'group STIT'.

branching-time account of STIT. Here, we present an alternative semantics for STIT based on the notion of STIT model with choice names (SCN) in which choices of agents are labelled with action identifiers. This can be seen as an explicit action-based version of the semantics for group STIT.[2] We show that this semantics is mathematically equivalent to the STIT semantics in terms of Kripke atemporal group STIT models introduced by [25] and, consequently, to the STIT semantics in terms of consequential models introduced by [30].

The main reason why we provide an alternative semantics based on STIT models with choice names is that this semantics is particularly suited to characterize the concepts of *interim* knowledge and *ex post* knowledge. Indeed, the semantic definition of an agent i's *interim* knowledge requires that, in all epistemic alternatives of agent i, i makes the *same* choice. This simply means that i's choice has the *same name* in all epistemic alternatives of agent i. Moreover, the semantic definition of an agent i's *ex post* knowledge requires that, in all epistemic alternatives of agent i, all agents make the *same* collective choice. This means that the collective choice of all agents has the *same name* in all epistemic alternatives of agent i.

In this paper we focus on Chellas's STIT operators, named after its proponent [11], and we take them as primitive. As pointed out by [28], the so-called deliberative STIT operators and the Chellas's STIT operators are inter-definable and they differ solely in their choice of primitive operators.

Before concluding, let us emphasize why we have decided to use STIT as a tool for our formal analysis of responsibility instead of some alternative logic of action (e.g., dynamic logic). The main reason is that STIT theory is compatible with both the idea of agent causation [12] (*i.e.*, the idea that an agent may be the cause of an event or state of affairs) and the indeterministic view of reality. These are essential elements for the analysis of responsibility attribution. On the one hand, we can ascribe to a certain agent i *active* responsibility for the realization of a state of affairs φ, only if agent i brings about φ thereby being the cause of the realization of φ. On the other hand, we can ascribe to i *passive* responsibility for the realization of φ, only if i could have done otherwise and prevented φ from being true.

2.1 Syntax

Assume a countable set of atomic propositions denoting facts $Atm = \{p, q, \ldots\}$ and a finite set of agents $Agt = \{1, \ldots, n\}$. $2^{Agt*} = 2^{Agt} \setminus \{\emptyset\}$ denotes the set of all non-empty sets of agents *alias* coalitions or groups.

The language $\mathcal{L}_{\text{E-GSTIT}}(Atm, Agt)$ of the logic E-GSTIT is the set of formulae defined by the following BNF:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid [J\ \mathbf{stit}]\varphi \mid \mathrm{K}_i^{\bullet\circ\circ}\varphi \mid \mathrm{K}_i^{\circ\bullet\circ}\varphi \mid \mathrm{K}_i^{\circ\circ\bullet}\varphi$$

where p ranges over Atm, J ranges over 2^{Agt} and i ranges over Agt. The other Boolean constructions $\top$, $\bot$, $\vee$, $\rightarrow$ and $\leftrightarrow$ are defined from $\neg$ and $\wedge$ in the standard way. As usual in modal logic, $\langle J\ \mathbf{stit}\rangle\varphi$ is an abbreviation of $\neg[J\ \mathbf{stit}]\neg\varphi$.

We call GSTIT (*Group* STIT) the fragment of E-GSTIT without epistemic operators of type $\mathrm{K}_i^{\bullet\circ\circ}\varphi$, $\mathrm{K}_i^{\circ\bullet\circ}$ and $\mathrm{K}_i^{\circ\circ\bullet}$, whose meaning is explained below. Specifically, the language

[2] According to van Benthem [43] the issue of defining an explicit action/strategy-based version of STIT is one of the fundamental problems in the area of logics of agency and game logics.

$\mathcal{L}_{\text{GSTIT}}(Atm, Agt)$ of the logic GSTIT is the set of formulae defined by the following BNF:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid [J\ \mathbf{stit}]\varphi$$

where p ranges over Atm and J ranges over 2^{Agt}.

Operators of type $[J\ \mathbf{stit}]$ are used to describe the effects of the action that has been chosen by J. In Belnap et al.'s STIT, an agent i's action is described in terms of the result that agent i brings about by her acting. For example, i's action of killing another agent j is described by the fact that i sees to it that j is dead. Likewise, in Horty's STIT with agents and groups, one can make a distinction between *individual actions* of agents and *joint actions* of groups of agents. The joint action of a group is described in terms of the result that the agents in the group bring about by acting together.

If $H \in 2^{Agt*}$, the construction $[H\ \mathbf{stit}]\varphi$ has to be read "group H sees to it that φ, no matter what the agents outside H do". Analogously, if H is a singleton $\{i\}$, the construction $[\{i\}\ \mathbf{stit}]\varphi$ has to be read "agent i sees to it that φ no matter what the other agents do". For notational convenience, we sometimes write $[i\ \mathbf{stit}]$ instead of $[\{i\}\ \mathbf{stit}]$.

$[\emptyset\ \mathbf{stit}]\varphi$ has to be read "φ is true regardless of what every agent does", or "φ is true no matter what the agents do", or simply 'φ is necessarily true'. The dual expression $\langle\emptyset\ \mathbf{stit}\rangle\varphi$ means "φ is possibly true". Note that the operators $\langle\emptyset\ \mathbf{stit}\rangle$ and $[J\ \mathbf{stit}]$ can be combined in order to express what agents and groups can do: $\langle\emptyset\ \mathbf{stit}\rangle[J\ \mathbf{stit}]\varphi$ means "J can see to it that φ, no matter what the agents outside J do". Conceptually speaking the difference between "see to it that" and "can see to it that" is that the former corresponds to the concept of *action* (*i.e.*, what an agent or group does) while the latter corresponds to the concept of *capability* (*i.e.*, what an agent or group can do).

Our logic E-GSTIT extends Horty's STIT logic with three different concepts and corresponding modal operators of knowledge. The operators $\mathrm{K}_i^{\bullet\circ\circ}$, $\mathrm{K}_i^{\circ\bullet\circ}$ and $\mathrm{K}_i^{\circ\circ\bullet}\varphi$ capture respectively the concepts of *ex ante* knowledge (*alias* knowledge before one's choice), *interim* knowledge (*alias* knowledge after one's choice and before being informed about the choices of others) and *ex post* knowledge (*alias* knowledge after public information about the choices of all agents is available). These three concepts of knowledge are taken from the vast literature on game theory (notably the reference article by Aumann and Dreze [3]), where the distinction between these three stages in differential information environments is widely accepted for games in normal form. This explicit borrowing is amply justified, because we may consider STIT models without time as nothing but game forms in the game theoretical sense. According to Aumann & Dreze, *ex ante* knowledge characterizes an agent's knowledge assuming that she has not made any decision yet, whereas *ex post* knowledge characterizes an agent's knowledge in a situation where every agent has made her choice and all choices are publicly disclosed. Finally, *interim* knowledge characterizes an agent's knowledge assuming that she has made her decision about which action to take, but might still be uncertain about the decisions of others. The formulae $\mathrm{K}_i^{\bullet\circ\circ}\varphi$, $\mathrm{K}_i^{\circ\bullet\circ}\varphi$ and $\mathrm{K}_i^{\circ\circ\bullet}\varphi$ have the following short readings: $\mathrm{K}_i^{\bullet\circ\circ}\varphi$ has to be read "the agent i has an *ex ante* knowledge that φ is true"; $\mathrm{K}_i^{\circ\bullet\circ}\varphi$ has to be read "the agent i has an *interim* knowledge that φ is true"; $\mathrm{K}_i^{\circ\circ\bullet}\varphi$ has to be read "the agent i has an *ex post* knowledge that φ is true". The corresponding dual operators are defined as $\widehat{\mathrm{K}}_i^x\varphi \stackrel{\text{def}}{=} \neg\mathrm{K}_i^x\neg\varphi$ with $x \in \{\bullet\circ\circ, \circ\bullet\circ, \circ\circ\bullet\}$.

2.2 Semantics

This section about semantics is organized in three subsections. In Section 2.2.1 we define the class of STIT models with choice names (SCNs) and prove that these structures are mathematically equivalent to Kripke atemporal group STIT models introduced by [25]. In Section 2.2.2 we add an extra epistemic component to SCNs by defining the class of *epistemic* STIT models with choice names (ESCNs). Finally, in Section 2.2.3 we provide an interpretation of E-GSTIT formulae over ESCNs.

2.2.1 STIT models with choice names

Assume a countable, possibly infinite, set of action terms $Act = \{a, b, \ldots\}$. If Act is infinite, then $card(Act) = \infty$, where $card(S)$ denotes the cardinality of an arbitrary set S. Similarly to game theory, we generalize actions to strategies by defining, for every coalition $H \in 2^{Agt*}$, the set Act_H of all total functions with domain H and codomain Act. Elements of Act_H are the joint actions of coalition H and are denoted by symbols $\alpha_H, \alpha'_H, \ldots$ Since α_H is a function $\alpha_H : H \longrightarrow Act$, this can be represented as a set of mappings $\{i \mapsto \alpha_H(i) | i \in H\}$. For example, $\alpha_{\{1,2,3\}} = \{1 \mapsto a, 2 \mapsto b, 3 \mapsto c\}$ is the joint action of coalition $\{1, 2, 3\}$ in which agent 1 chooses action a, agent 2 chooses action b and agent 3 chooses action c.

The basic notion here is the notion of STIT model with choice names that is nothing but a collection of binary relations on a set of worlds and a set of functions specifying the action chosen by an agent at a certain world. Given an arbitrary binary relation $\mathcal{R}$ on a set of elements W, let $\mathcal{R}(w) = \{v \in W | w\mathcal{R}v\}$.

Definition 1 (STIT model with choice names (SCN)) *A STIT model with choice names (SCN) is a tuple $M = \langle W, \{\mathcal{A}_i\}_{i \in Agt}, \{\mathcal{R}_J\}_{J \subseteq Agt}, \mathcal{V}\rangle$ where:*

- *W is a nonempty set of possible worlds;*
- *for every $i \in Agt$, $\mathcal{A}_i : W \longrightarrow Act$ is a total function mapping worlds into actions;*
- *$\mathcal{R}_\emptyset$ is an equivalence relation between worlds in W satisfying the following condition:*

  (C1) *for all $w, u_1, \ldots, u_n \in W$ and for all $a_1, \ldots, a_n \in Act$, if $u_1, \ldots, u_n \in \mathcal{R}_\emptyset(w)$ and $\mathcal{A}_1(u_1) = a_1, \ldots, \mathcal{A}_n(u_n) = a_n$ then there exists $u \in W$ such that $u \in \mathcal{R}_\emptyset(w)$ and $\mathcal{A}_{Agt}(u) = \{1 \mapsto a_1, \ldots, n \mapsto a_n\}$;*

- *every $\mathcal{R}_H$ with $H \in 2^{Agt*}$ is a binary relation between worlds in W such that:*

  (C2) *$\mathcal{R}_H = \{(w, v) | w\mathcal{R}_\emptyset v$ and $\mathcal{A}_H(w) = \mathcal{A}_H(v)\}$;*

- *$\mathcal{V} : Atm \longrightarrow 2^W$ is a valuation function for atomic formulae;*

where for every $H \in 2^{Agt*}$, $\mathcal{A}_H$ is a total function $\mathcal{A}_H : W \longrightarrow Act_H$ such that $\mathcal{A}_H(w) = \alpha_H$ if and only if $\mathcal{A}_i(w) = \alpha_H(i)$ for all $i \in H$.

Let us discuss the preceding notion of SCN in detail. Each function $\mathcal{A}_i$ specifies the action chosen by agent i at given world. Specifically, $\mathcal{A}_i(w)$ is the action chosen by the agent i at world w.

$\mathcal{R}_\emptyset(w)$ is the set of outcomes of the agents' choices that are possible at world w: if $v \in \mathcal{R}_\emptyset(w)$, then, at world w, the agents can choose a joint action whose outcome is v.

For every world w and for every coalition $H \in 2^{Agt*}$, the set $\mathcal{R}_H(w)$ identifies coalition H's *actual* choice at w, that is to say, the set of worlds that can be reached by coalition H's actual choice at w. Note that, if $card(H) = 1$, $\mathcal{R}_H(w)$ is an *individual* choice, whereas if $card(H) > 1$, $\mathcal{R}_H(w)$ is a *collective* choice. Hence, if $v \in \mathcal{R}_H(w)$ then v is a world that is *admitted* by coalition H's actual choice at world w.

Note that, if v is admitted by coalition H's actual choice at world w (*i.e.*, $v \in \mathcal{R}_H(w)$), this means that, given what the agents in H have chosen at w, there exists a choice of the agents in $Agt \setminus H$ such that, if the agents in $Agt \setminus H$ made such choice, v would be a possible outcome of the collective choice of all agents. Consequently, $v \in \mathcal{R}_{Agt\setminus H}(w)$ means that, given what the agents in H have chosen at w, there exists a choice of the agents in $Agt \setminus H$ such that, if the agents in $Agt \setminus H$ made such choice, v would be the actual outcome of the collective choice of all agents.

Condition **C1** corresponds to a basic game theory assumption according to which, if agents can choose their individual actions separately, they can choose them jointly as well.

Condition **C2** specifies the relationship between the set of possible outcomes and the set of worlds that can be reached by the actual choice of a coalition. According to this constraint, the set of worlds that can be reached by the actual choice of a coalition at w is the set of possible outcomes of the joint action that the coalition chooses at w.

We can show that SCNs are mathematically equivalent to Kripke atemporal group STIT models, as defined in [25, 35], that in turn are equivalent to consequentialist models in the sense of [30].[3] The interesting aspect of proving the equivalence between SCNs and Kripke atemporal group STIT models is that this result clarifies the relationship between the standard semantics of STIT and the representation of social interaction in game theory. Indeed, STIT models with choice names (SCNs) are very similar to games in normal form.

The following definition introduces Kripke atemporal group STIT models, or Kripke group STIT models for short.

Definition 2 (Kripke group STIT model [25, 35]) *A Kripke group STIT model is a tuple* $M = \langle W, \{\mathcal{R}_J\}_{J\subseteq Agt}, \mathcal{V}\rangle$ *where:*

- $\mathcal{R}_\emptyset$ *and every* $\mathcal{R}_H$ *with* $H \in 2^{Agt*}$ *are equivalence relations between worlds in* W *satisfying the following conditions:*

  (C1*) $\mathcal{R}_H \subseteq \mathcal{R}_\emptyset$,

  (C2*) $\mathcal{R}_H = \bigcap_{i\in H} \mathcal{R}_{\{i\}}$,

  (C3*) *for all* $w, u_1, \ldots, u_n \in W$, *if* $u_1, \ldots, u_n \in \mathcal{R}_\emptyset(w)$ *then* $\bigcap_{1\leq i\leq n} \mathcal{R}_{\{i\}}(u_i) \neq \emptyset$;

- $\mathcal{V} : Atm \longrightarrow 2^W$ *is a valuation function for atomic formulae.*

Condition **C1*** states that the set of worlds that can be reached by the actual choice of a coalition is a subset of the set of possible outcomes. In other words, a coalition can only choose among possible outcomes. According to Condition **C2***, the set of worlds that can be reached by the choice of a coalition is equal to the pointwise intersection of the sets of worlds that can

[3] As shown by [25], the semantics of group STIT in terms of Kripke atemporal group STIT models is equivalent to the semantics in terms of S5^n product models [15]. It is also worth noting that this result has been independently proved by Gerbrandy [16] who has shown that the most general instance of the logic of propositional control can also be interpreted over S5^n product models. The logic of propositional control studied by Gerbrandy and atemporal group STIT are indeed the same logic.

be reached by the individual choices of the agents in the coalition. That is, the collective choice of a coalition is made up of the individual choices of the individual agents in the coalition and nothing else. Condition **C3*** expresses the so-called *assumption of independence of agents*. Intuitively, it means that agents can never be deprived of choices due to the choices made by other agents.

The previous definition of Kripke group STIT model does not impose any restriction on the number of available choices for an agent in a given state. Indeed, the STIT semantics allows an agent to have an infinite number of choices. The following definition introduces the concept of Kripke group STIT model with bounded choices, *i.e.*, the class of Kripke group STIT models in which the number of available choices for an agent is bounded by some $n \in \mathbb{N} \cup \{\infty\}$.

Definition 3 (Kripke group STIT model with bounded choices) *Let $M = \langle W, \{\mathcal{R}_J\}_{J \subseteq Agt}, \mathcal{V}\rangle$ be a Kripke group STIT model and let $\mathbf{P}$ be the partition of W induced by the equivalence relation $\mathcal{R}_\emptyset$. Moreover, let $n \in \mathbb{N} \cup \{\infty\}$. M is said to be a Kripke group STIT model whose number of choices is bounded by n if and only if, for all $i \in Agt$ and for all $X \in \mathbf{P}$ the number of parts in the partition of X induced by the relation $\mathcal{R}_{\{i\}}$ is equal to or less than n.*

Note that the preceding definition also covers the case of an unbounded model M in which an agent has an infinite number of available choices, *i.e.*, there exists $i \in Agt$ and $X \in \mathbf{P}$ such that the number of parts in the partition of X induced by the relation $\mathcal{R}_{\{i\}}$ is infinite (∞).

The following representation theorem establishes the formal connection between STIT models with choice names and Kripke group STIT models.

Theorem 1 *Let $card(Act) = n$ with $n \in \mathbb{N} \cup \{\infty\}$. Then:*

- *if $\langle W, \{\mathcal{R}_J\}_{J \subseteq Agt}, \{\mathcal{A}_i\}_{i \in Agt}, \mathcal{V}\rangle$ is a SCN then $\langle W, \{\mathcal{R}_J\}_{J \subseteq Agt}, \mathcal{V}\rangle$ is a Kripke group STIT model whose number of choices is bounded by n;*
- *if $\langle W, \{\mathcal{R}_J\}_{J \subseteq Agt}, \mathcal{V}\rangle$ is a Kripke group STIT model whose number of choices is bounded by n then there exists a set of functions $\{\mathcal{A}_i\}_{i \in Agt}$ such that $\langle W, \{\mathcal{R}_J\}_{J \subseteq Agt}, \{\mathcal{A}_i\}_{i \in Agt}, \mathcal{V}\rangle$ is a SCN.*

PROOF. Let us prove the first item by supposing that $card(Act) = n$ and by taking a SCN $M = \langle W, \{\mathcal{R}_J\}_{J \subseteq Agt}, \{\mathcal{A}_i\}_{i \in Agt}, \mathcal{V}\rangle$. We are going to show that $M' = \langle W, \{\mathcal{R}_J\}_{J \subseteq Agt}, \mathcal{V}\rangle$ is a Kripke group STIT model. Clearly, every relation $\mathcal{R}_J$ is an equivalence relation and M' satisfies **C1***. As to **C2***, by Condition **C2**, we have $\mathcal{R}_J = \{(w,v) | w\mathcal{R}_\emptyset v \text{ and } \mathcal{A}_J(w) = \mathcal{A}_J(v)\} = \{(w,v) | w\mathcal{R}_\emptyset v \text{ and } \mathcal{A}_i(w) = \mathcal{A}_i(v) \text{ for all } i \in J\} = \bigcap_{i \in J}\{(w,v) | w\mathcal{R}_\emptyset v \text{ and } \mathcal{A}_i(w) = \mathcal{A}_i(v)\} = \bigcap_{i \in J} \mathcal{R}_{\{i\}}$. Let us prove that model M' satisfies Condition **C3***, let us suppose that $u_1 \in \mathcal{R}_\emptyset(w), \ldots, u_n \in \mathcal{R}_\emptyset(w)$. Then, because every function $\mathcal{A}_i$ is total, we have that there are $a_1, \ldots, a_n \in Act$ such that $\mathcal{A}_1(u_1) = a_1, \ldots, \mathcal{A}_n(u_n) = a_n$. Hence, by Condition **C1**, there are $a_1, \ldots, a_n \in Act$ and $u \in W$ such that $\mathcal{A}_1(u_1) = a_1, \ldots, \mathcal{A}_n(u_n) = a_n$, $u \in \mathcal{R}_\emptyset(w)$ and $\mathcal{A}_{Agt}(u) = \{1 \mapsto a_1, \ldots, n \mapsto a_n\}$. By definition of $\mathcal{R}_{\{i\}}$, it follows that there is $u \in W$ such that $u_1\mathcal{R}_{\{1\}}u, \ldots, u_n\mathcal{R}_{\{n\}}u$. Hence, $\bigcap_{1 \leq i \leq n} \mathcal{R}_{\{i\}}(u_i) \neq \emptyset$. By the fact that $card(Act) = n$, it is easy to verify that for every $X \in \mathbf{P}$ the number of parts in the partition of X induced by the relation $\mathcal{R}_{\{i\}}$ is equal to or less than n. Thus, M' is a Kripke group STIT model in which the number of choices is bounded by n.

Let us prove the second item by supposing that $card(Act) = n$. Let us take an arbitrary Kripke group STIT model $M = \langle W, \{\mathcal{R}_J\}_{J \subseteq Agt}, \mathcal{V}\rangle$ whose number of choices is bounded

by n. Let $\mathbf{P}$ be the partition of the set of worlds W induced by the equivalence relation $\mathcal{R}_\emptyset$. For each agent $i \in Agt$, we define $\mathcal{A}_i$ to be a total function $\mathcal{A}_i : W \longrightarrow Act$ satisfying the following condition:

(*) for all $X \in \mathbf{P}$ and for all $w, v \in X$, $\mathcal{A}_i(w) = \mathcal{A}_i(v)$ if and only if $w\mathcal{R}_{\{i\}}v$.

Such a function $\mathcal{A}_i$ is guaranteed to exist by the fact that $card(Act) = n$ and the fact that for every $X \in \mathbf{P}$ the number of parts in the partition of X induced by the relation $\mathcal{R}_{\{i\}}$ is equal to or less than n. Moreover, it is a routine task to verify that the model $M' = \langle W, \{\mathcal{R}_J\}_{J\subseteq Agt}, \{\mathcal{A}_i\}_{i\in Agt}, \mathcal{V}\rangle$ so generated is a SCN. Indeed, M' clearly satisfies Condition **C2**. We are going to prove that it satisfies Condition **C1**. Let us suppose that $u_1, \ldots, u_n \in \mathcal{R}_\emptyset(w)$ and $\mathcal{A}_1(u_1) = a_1, \ldots, \mathcal{A}_n(u_n) = a_n$. By Condition **C3*** it follows that $\bigcap_{1\leq i\leq n} \mathcal{R}_{\{i\}}(u_i) \neq \emptyset$. Hence, by the preceding Condition (*),

$$\bigcap_{1\leq i\leq n}\{u \in W | w\mathcal{R}_\emptyset u \text{ and } \mathcal{A}_i(u) = \mathcal{A}_i(u_i) = a_i\} \neq \emptyset.$$

The latter implies that there exists $u \in W$ such that $u \in \mathcal{R}_\emptyset(w)$ and $\mathcal{A}_{Agt}(u) = \{1 \mapsto a_1, \ldots, n \mapsto a_n\}$. ■

From Theorem 1 it follows that if $card(Act) = \infty$ then, for every Kripke group STIT model $\langle W, \{\mathcal{R}_J\}_{J\subseteq Agt}, \mathcal{V}\rangle$, there exists a set of functions $\{\mathcal{A}_i\}_{i\in Agt}$ such that $\langle W, \{\mathcal{R}_J\}_{J\subseteq Agt}, \{\mathcal{A}_i\}_{i\in Agt}, \mathcal{V}\rangle$ is a SCN.

2.2.2 Epistemic STIT models with choice names

The following definition introduces the concept of epistemic STIT model with choice names (ESCN), namely a SCN supplemented with an epistemic component.

Definition 4 (Epistemic STIT model with choice names (ESCN)) *An epistemic STIT model with choice names (ESCN) is a tuple*

$$M = \langle W, \{\mathcal{A}_i\}_{i\in Agt}, \{\mathcal{R}_J\}_{J\subseteq Agt}, \{\mathcal{E}_i^{\bullet\circ\circ}, \mathcal{E}_i^{\circ\bullet\circ}, \mathcal{E}_i^{\circ\circ\bullet}\}_{i\in Agt}, \mathcal{V}\rangle$$

where:

- $M = \langle W, \{\mathcal{A}_i\}_{i\in Agt}, \{\mathcal{R}_J\}_{J\subseteq Agt}, \mathcal{V}\rangle$ *is a SCN;*
- *all* $\mathcal{E}_i^{\bullet\circ\circ}$ *are equivalence relations between worlds in* W*;*
- *all* $\mathcal{E}_i^{\circ\bullet\circ}$ *and* $\mathcal{E}_i^{\circ\circ\bullet}$ *are binary relations between worlds in* W *such that:*

  (C3) $\mathcal{E}_i^{\circ\bullet\circ} = \{(w, v) | w\mathcal{E}_i^{\bullet\circ\circ}v$ *and* $\mathcal{A}_i(w) = \mathcal{A}_i(v)\}$,

  (C4) $\mathcal{E}_i^{\circ\circ\bullet} = \{(w, v) | w\mathcal{E}_i^{\bullet\circ\circ}v$ *and* $\mathcal{A}_{Agt}(w) = \mathcal{A}_{Agt}(v)\}$.

We suppose that each agent has three different accessibility relations $\mathcal{E}_i^{\bullet\circ\circ}$, $\mathcal{E}_i^{\circ\bullet\circ}$ and $\mathcal{E}_i^{\circ\circ\bullet}$ corresponding to *ex ante*, *interim* and *ex post* knowledge. In particular, $\mathcal{E}_i^{\bullet\circ\circ}(w)$ is the set of worlds that agent i considers possible before making her choice and before being informed about the choices of other agents. We call it i's *ex ante information set* at w. $\mathcal{E}_i^{\circ\bullet\circ}(w)$ is the set of worlds that agent i considers possible after making her choice but before being informed about the choices of other agents. We call it i's *interim information set* at w. Finally, $\mathcal{E}_i^{\circ\circ\bullet}(w)$ is the set of worlds that agent i considers possible after having made her choice and after being informed about the choices of others. We call it i's *ex post information set* at w.

Conditions **C3** and **C4** characterize the relationships between the three types of information sets. According to Condition **C3**, agent i's *interim information set* consists in restricting agent

i's *ex ante information set* to the worlds which are consistent with i's current choice. According to Condition **C4**, agent i's *ex post information set* is the result of restricting agent i's *ex ante information set* to the worlds which are consistent with the current choices of all agents. Note that Conditions **C3** and **C4** ensure that $\mathcal{E}_i^{\circ\bullet\circ}$ and $\mathcal{E}_i^{\circ\circ\bullet}$ are also equivalence relations.

Remark. In some cases, it would be reasonable to suppose that $\mathcal{R}_{Agt\setminus\{i\}} \subseteq \mathcal{E}_i^{\bullet\circ\circ}$ for all $i \in Agt$. This condition means that, before making her choice, an agent considers possible all her available choices. However, we do not impose this constraint because we want to allow situations in which an agent excludes some of her available decision options *a priori*. For example, a person may decide to kill herself by jumping from a balcony (*i.e.*, the action of jumping from the balcony is an available decision option), even though she considers this choice *epistemically* impossible.

2.2.3 Truth conditions of E-GSTIT formulae

A formula φ of the logic E-GSTIT is evaluated with respect to a given ESCN $M = \langle W, \{\mathcal{A}_i\}_{i\in Agt}, \{\mathcal{R}_J\}_{J\subseteq Agt}, \{\mathcal{E}_i^{\bullet\circ\circ}, \mathcal{E}_i^{\circ\bullet\circ}, \mathcal{E}_i^{\circ\circ\bullet}\}_{i\in Agt}, \mathcal{V}\rangle$ and a world w in M. We write $M, w \models \varphi$ to mean that φ is true at world w in M. The truth conditions of E-GSTIT formulae are then defined as follows.

Definition 5 (Truth conditions of E-GSTIT formulae)
Let M be an $ESCN$ and w a world in M:

$$\begin{aligned}
M, w \models p &\iff w \in \mathcal{V}(p)\\
M, w \models \neg\varphi &\iff M, w \not\models \neg\varphi\\
M, w \models \varphi \wedge \psi &\iff M, w \models \varphi \text{ AND } M, w \models \psi\\
M, w \models [J\ \mathbf{stit}]\varphi &\iff \forall v \in \mathcal{R}_J(w) : M, v \models \varphi\\
M, w \models \mathrm{K}_i^{\bullet\circ\circ}\varphi &\iff \forall v \in \mathcal{E}_i^{\bullet\circ\circ}(w) : M, v \models \varphi\\
M, w \models \mathrm{K}_i^{\circ\bullet\circ}\varphi &\iff \forall v \in \mathcal{E}_i^{\circ\bullet\circ}(w) : M, v \models \varphi\\
M, w \models \mathrm{K}_i^{\circ\circ\bullet}\varphi &\iff \forall v \in \mathcal{E}_i^{\circ\circ\bullet}(w) : M, v \models \varphi
\end{aligned}$$

For any formula φ of the language $\mathcal{L}_{\text{GSTIT}}(Atm, Agt)$, we write $\models_{\text{GSTIT}} \varphi$ if φ is GSTIT *valid*, that is, if φ is true in all SCNs (*i.e.*, for all ECNs M and for all worlds w in M, we have $M, w \models \varphi$). We say that φ is GSTIT *satisfiable* if $\neg\varphi$ is not GSTIT valid. Moreover, for any formula φ of the language $\mathcal{L}_{\text{E-GSTIT}}(Atm, Agt)$, we write $\models_{\text{E-GSTIT}} \varphi$ if φ is E-GSTIT *valid*, that is, if φ is true in all ESCNs (*i.e.*, for all ESCNs M and for all worlds w in M, we have $M, w \models \varphi$). We say that φ is E-GSTIT *satisfiable* if $\neg\varphi$ is not E-GSTIT valid.

The following example is given in order to illustrate the notion of epistemic STIT model with choice names (ESCN) given in Definition 4 and the truth conditions of E-GSTIT formulae.

Example 1 *Let us suppose that $Agt = \{1, 2\}$. Both agents 1 and 2 are writing an article together and each of them can contribute to the joint activity by devoting either a small (action S), a medium (action M) or a large (action L) quantity of work to the paper. The result of the joint action resulting from both the action of agent 1 and that of agent 2 can be either that the paper is accepted for publication (p) or that it is not ($\neg p$). The agents have some uncertainty*

about the effects of their joint actions. For instance, agent 1 *does not know whether, if they both devote a medium quantity of work to the paper, it will be accepted.*

This situation is represented by the ESCN in Figure 1 in which agent 1*'s three types of information sets are drawn in the three different subfigures 1a, 1b and 1c: agent* 1*'s* ex ante information set *(Figure 1a), agent* 1*'s* interim information set *(Figure 1b) and agent* 1*'s* ex post information set *(Figure 1c). Let us assume that the actual world is the world* w *in the left grid.*

In the actual world w *agents* 1 *and* 2 *can ensure that the paper will be accepted if and only if each of them devotes at least a medium quantity of work. Indeed, for all* $v \in \mathcal{R}_\emptyset(w)$ *we have* $M, v \models p$ *if and only if*

$$\mathcal{A}_{\{1,2\}}(v) \in \{\{1 \mapsto L, 2 \mapsto L\}, \{1 \mapsto L, 2 \mapsto M\}, \{1 \mapsto M, 2 \mapsto L\}, \{1 \mapsto M, 2 \mapsto M\}\}.$$

Furthermore, in the actual world w*, before making her choice, agent* 1 *has some uncertainty about the effect of the joint action* $\{1 \mapsto M, 2 \mapsto M\}$*. Indeed, at world* w *in Figure 1a agent* 1 *envisages two possible scenarios: a scenario in which they can ensure that the paper will be accepted by both devoting a medium quantity of work (the left grid), and a scenario in which they cannot ensure so (the right grid). Formally, in Figure 1a we have: (i) there are* $v_1 \in \mathcal{E}_1^{\bullet\circ\circ}(w)$ *and* $u_1 \in \mathcal{R}_\emptyset(v_1)$ *such that* $\mathcal{A}_{\{1,2\}}(u_1) = \{1 \mapsto M, 2 \mapsto M\}$ *and* $M, u_1 \models [\{1,2\}\ \mathbf{stit}]p$*; and (ii) there is* $v_2 \in \mathcal{E}_1^{\bullet\circ\circ}(w)$ *such that for all* $u_2 \in \mathcal{R}_\emptyset(v_2)$ *if* $\mathcal{A}_{\{1,2\}}(u_2) = \{1 \mapsto M, 2 \mapsto M\}$ *then* $M, u_2 \models \neg[\{1,2\}\ \mathbf{stit}]p$*.*

Note that, before making her choice, 1 *does not exclude any of her available choices, in the sense that the three actions* L*,* M *and* S *are all within agent* 1*'s ex ante information set. This means that, for every action in* $\{L, M, S\}$*, before making her choice* 1 *envisages a world in which she chooses such action (see Figure 1a). Moreover, after having made her choice and before learning the choice of agent* 2*,* 1 *has a maximal uncertainty about it,* i.e.*,* 1 *does not know whether* 2 *is going to choose action* S*, action* M *or action* L *(see Figure 1b).*

Note also that a difference between 1*'s* ex ante *knowledge and* 1*'s* interim *knowledge at* w *is the following: before choosing what to do herself,* 1 *is uncertain whether they are going to see to it that the paper is rejected (Figure 1a); whereas, after choosing what to do (action* S*),* 1 *knows that they are going to see to it that the paper is rejected (Figure 1b). That is, in terms of E-GSTIT formulae we have:*

$$M, w \models \widehat{\mathrm{K}}_1^{\bullet\circ\circ}[\{1,2\}\ \mathbf{stit}]\neg p \wedge \widehat{\mathrm{K}}_1^{\bullet\circ\circ}\neg[\{1,2\}\ \mathbf{stit}]\neg p \wedge \mathrm{K}_1^{\circ\bullet\circ}[\{1,2\}\ \mathbf{stit}]\neg p$$

Before concluding this section let us consider some valid principles of the logic E-GSTIT. For all $x \in \{\bullet\circ\circ, \circ\bullet\circ, \circ\circ\bullet\}$ we have:

$$\models_{\text{E-GSTIT}} (\mathrm{K}_i^x\varphi \wedge \mathrm{K}_i^x\psi) \to \mathrm{K}_i^x(\varphi \wedge \psi) \tag{1}$$

$$\models_{\text{E-GSTIT}} \mathrm{K}_i^x\varphi \to \varphi \tag{2}$$

$$\models_{\text{E-GSTIT}} \mathrm{K}_i^x\varphi \to \mathrm{K}_i^x\mathrm{K}_i^x\varphi \tag{3}$$

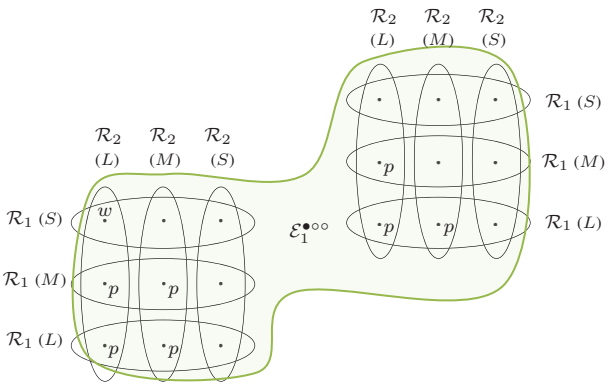
$$\models_{\text{E-GSTIT}} \neg\mathrm{K}_i^x\varphi \to \mathrm{K}_i^x\neg\mathrm{K}_i^x\varphi \tag{4}$$

$$\text{If } \models_{\text{E-GSTIT}} \varphi \text{ then } \models_{\text{E-GSTIT}} \mathrm{K}_i^x\varphi \tag{5}$$

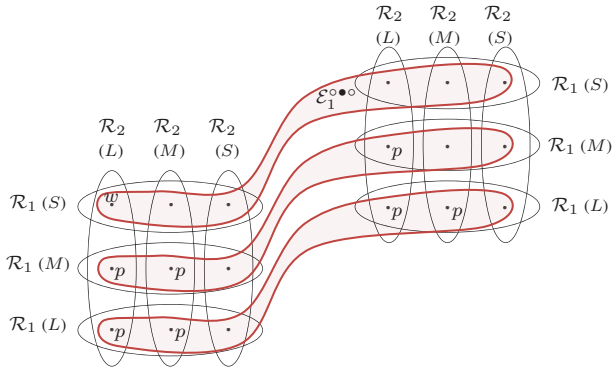
The previous validities (1)-(4) together with the rule of inference (5) highlight that *ex ante*, *ex interim*, and *ex post* knowledge operators are all S5 normal modal operators.

$$\models_{\text{E-GSTIT}} \mathrm{K}_i^{\bullet\circ\circ}\varphi \to \mathrm{K}_i^{\circ\bullet\circ}\varphi \tag{6}$$

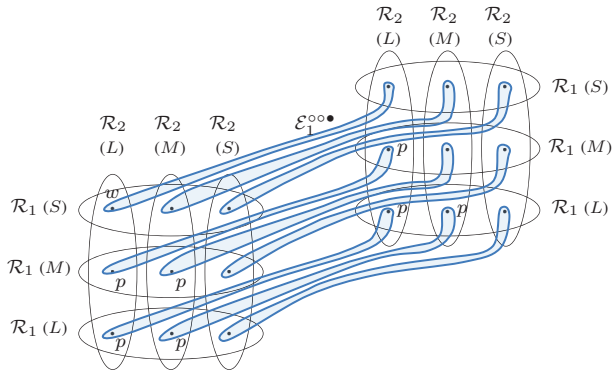
$$\models_{\text{E-GSTIT}} \mathrm{K}_i^{\circ\bullet\circ}\varphi \to \mathrm{K}_i^{\circ\circ\bullet}\varphi \tag{7}$$



(a) Ex ante information set



(b) Interim information set



(c) Ex post information set

Figure 1: Example of two agents writing a paper both together

Validities (6) and (7) characterize the two basic relationships between *ex ante*, *interim* and *ex post* knowledge: *first,* an agent knows more after the choices of all agents have became public, than after having made her choice but before being informed about the choices of other agents; and *second,* an agent knows more after having made her choice but before being informed about the choices of other agents, than before having made her choice and before being informed about others' choices.

3 Responsibility

The logic E-GSTIT presented in section 2 constitutes the departure point for a theoretical model of (*agentive*) responsibility that will pave the way to construct a formal model of attribution emotions in Section 4. The term 'responsibility' has more than one meaning and is commonly used in many senses. Therefore, let us devote this section to clarify some conceptual differences among the different usages of the term, mainly: *causal* and *agentive* responsibility.[4]

Perhaps the root notion of responsibility relates to the use of the term to indicate mere *causal* relation (*e.g.* "The fire was responsible for the damages in the building"). When we use the term 'responsible' this way, we solely point out the causal relationship[5] between both elements; and thus, we may attribute responsibility without imputing intentionality to the causal origin of a given outcome. This is also the case of unintentional actions of an agent that lead to a given outcome (*e.g.* when an agent accidentally drops something, causing damages to another object); then the agent is considered causally responsible.

However, an agent's being causally responsible for an outcome is necessary but not sufficient for agentive responsibility. The reference to the conditions for holding an individual *agentive responsible* has been a constant in the philosophical construction of the concept of responsibility. In his *Nicomachean Ethics* [2, 1113b], Aristotle constructs the first explicit concept of agentive responsibility. According to this Aristotelian account, a given agent i is said to be agentive responsible for an event or state of affairs φ (*e.g.*, agent j is dead) if and only if agent i's action that causes the state is voluntary. Moreover, Aristotle identifies two distinctive features of a voluntary action:

- *Control condition*: the action must be the consequence of the agent's actual choice in the sense that, by making a given choice, the agent causes φ to be true (*e.g.*, *by choosing* to shoot with the gun, agent i *causes* agent j's death);
- *Epistemic condition*: the agent must be fully aware of what it is she is doing or bringing about (in the previous example, agent i *knows* that, by choosing to shoot the gun, she causes agent j's death).

Regarding both the *causal* and *agentive* responsibility, one of the possible misconceptions of the concept of responsibility is derived from the fact of narrowing it down to its active version; that is, to solely focus on the consequences that the agent's actions have for other agents. However, in order to have a broader account of the concept, we must also include the consequences of the actions that the agent could have chosen to perform and did not (see

[4] This distinction is widely accepted, although the terminology used varies. Cf., *e.g.*, [46].

[5] The concept of causation has deeper implications in many areas, notably for example, regarding legal consequences (cf. for instance, Hart and Honoré [24].). An analysis of this sense of responsibility would require a further inquiry into the notion of causation, which is out of the scope of this article. Cf. Hart, [23, Chapter 9], for a further reading on causal responsibility and its (legal) implications.

[19, p. 228] for instance). Hence, our analysis of responsibility also includes the *active* (the agent's ability to see to it that φ) and *passive* (the agent's ability to prevent φ from happening) dimensions of responsibility.

The four types of responsibility derived from the combination of the *causal-agentive* and *active-passive* dimensions are represented in the following Table 1.

Table 1: The different types of responsibility

		Responsibility	
		Active Responsibility	Passive Responsibility
Knowledge	Knowledge	Active Agentive Responsibility	Passive Agentive Responsibility
	No knowledge	Active Causal Responsibility	Passive Causal Responsibility

In our framework, an agent is said to be causally responsible for an event, if she is the actual cause of the event in the sense that, either she sees to it that φ is true (*active causal responsibility*) or she could have prevented φ from being true (*passive causal responsibility*). Following the Aristotelian view on responsibility,[6] we claim that a mere causal connection between the agent and a given outcome is not enough for attributing her agentive responsibility for such outcome. In order to be agentive responsible for an event φ, the agent must know either that her current choice makes φ true (*active agentive responsibility*) or that she could have prevented φ from being true (*passive agentive responsibility*).

All four concepts of Table 1 are expressible in the logic E-GSTIT. Let us first consider *active causal responsibility*. This concept is expressed by the so-called 'deliberative' STIT operator $[i\ \mathbf{dstit}]$ [28] which is definable in the following way:

$$[i\ \mathbf{dstit}]\varphi \stackrel{\text{def}}{=} [i\ \mathbf{stit}]\varphi \wedge \langle \emptyset\ \mathbf{stit}\rangle \neg\varphi$$

The construction $[i\ \mathbf{dstit}]\varphi$ can be read "agent i is causally responsible for bringing about φ". The negative condition $\langle \emptyset\ \mathbf{stit}\rangle \neg\varphi$ is given to prevent an agent from being causally responsible for φ, when φ is something inevitable (in the sense that it is true regardless of what every agent does). Consequently, *active agentive responsibility* is defined by the abbreviation:

$$\mathrm{Resp}^{+}(i,\varphi) \stackrel{\text{def}}{=} \mathrm{K}_i^{\circ\bullet\circ}[i\ \mathbf{dstit}]\varphi$$

Agent i is actively agentive responsible for bringing about φ (denoted by $\mathrm{Resp}^{+}(i,\varphi)$) if and only if i knows that her current choice makes φ true. In other words, i *knowingly* sees to it that φ.[7] For example, agent 1 can be said to be actively agentive responsible for killing agent 2 (*i.e.*, $\mathrm{Resp}^{+}(1,dead_2)$) if and only if agent 1 knows that her current choice will lead to agent 2's death and that agent 2's death is not something inevitable (*i.e.*, $\mathrm{K}_1^{\circ\bullet\circ}[i\ \mathbf{dstit}]\varphi$). We use

[6] Modern theories of responsibility are based on Aristotle's theory. See, *e.g.*, Pettit's model of responsibility attribution [38], based on three conditions: first, *value relevance*, the agent faces a morally significant choice; second, *value judgment*, the agent is in a position to see what is at stake; and third, *value sensitivity*, the choice was within the domain of the agent's will or control.

[7] A similar concept of *knowingly doing* is studied by [9].

the *interim* knowledge operator $\mathrm{K}_i^{\circ\bullet\circ}$ (instead of $\mathrm{K}_i^{\bullet\circ\circ}$ or $\mathrm{K}_i^{\circ\circ\bullet}$) in the preceding definition because, in order to be morally responsible for something, an agent has to be aware of the potential effects of her choice after having chosen it but before knowing the choices of other agents.

As for *passive causal responsibility*, we say that an agent i is passively causal responsible for letting φ be true if and only if "agent i could have prevented φ from being true", denoted by $\mathrm{CHP}(i,\varphi)$. In E-GSTIT logic the construction $\mathrm{CHP}(i,\varphi)$ can be decomposed by assuming that agent i could have prevented φ from being true if and only if: (1) φ is true in the actual world (*i.e.*, φ) and (2) given what the others have chosen to do, if i had chosen differently then φ would have necessarily been false (*i.e.*, $\langle Agt \setminus \{i\}\ \mathbf{stit}\rangle[Agt\ \mathbf{stit}]\neg\varphi$).[8] This notion of *passive causal responsibility* has already appeared in some of our previous works [35].

$$\mathrm{CHP}(i,\varphi) \stackrel{\text{def}}{=} \varphi \wedge \langle Agt \setminus \{i\}\ \mathbf{stit}\rangle[Agt\ \mathbf{stit}]\neg\varphi$$

The following is the semantic counterpart of the construction $\mathrm{CHP}(i,\varphi)$. We have that $M, w \models \mathrm{CHP}(i,\varphi)$ if and only if, $M, w \models \varphi$ and there is $v \in \mathcal{R}_{Agt\setminus\{i\}}(w)$ such that $M, v \models [Agt\ \mathbf{stit}]\neg\varphi$. That is, at world w of model M, agent i could have prevented φ from being true if and only if, φ is true at w and, given what the other agents in $Agt \setminus \{i\}$ have chosen at w, there exists an action of agent i such that, if agent i did choose this action, the actual outcome of the joint action of all agents would necessarily be a state in which φ is false.

In order to illustrate the previous definition of $\mathrm{CHP}(i,\varphi)$ let us go back to the example of Section 2.2.

Example 2 *At world w in the model M of Figure 1 we have that agent 1 could have prevented the paper from being rejected. That is:*

$$M, w \models \mathrm{CHP}(1,\neg p)$$

Indeed, we have that at world w: (1) p is false and (2) given what agent 2 has chosen to do at w (action L), there exists an action of agent 1 (action L or action M) such that, if agent 1 did choose this action, the actual outcome of the joint action of agents 1 and 2 would necessarily be a state in which p is true. That is:

$$M, w \models \neg p \wedge \langle\{2\}\ \mathbf{stit}\rangle[\{1,2\}\ \mathbf{stit}]p$$

As for *passive agentive responsibility*, we say that an agent i is passively agentive responsible for letting φ be true, denoted by $\mathrm{Resp}^{-}(i,\varphi)$, if and only if: (1) i knows that her current choice leads to φ (*i.e.*, $\mathrm{K}_i^{\circ\bullet\circ}\varphi$) and (2) given what the others have chosen to do, if i had chosen differently then, after having made her choice, she would have necessarily known that $\neg\varphi$ (*i.e.*, $\langle Agt \setminus \{i\}\ \mathbf{stit}\rangle[Agt\ \mathbf{stit}]\mathrm{K}_i^{\circ\bullet\circ}\neg\varphi$). That is:

$$\mathrm{Resp}^{-}(i,\varphi) \stackrel{\text{def}}{=} \mathrm{K}_i^{\circ\bullet\circ}\varphi \wedge \langle Agt \setminus \{i\}\ \mathbf{stit}\rangle[Agt\ \mathbf{stit}]\mathrm{K}_i^{\circ\bullet\circ}\neg\varphi$$

Example 3 *In the example of Figure 1, at world w agent 1 is* not *passive agentive responsible for letting the paper be rejected. Indeed, at world w the second condition of the preceding definition of passive agentive responsibility does not hold. In particular, we have*

$$M, w \models \neg\langle\{2\}\ \mathbf{stit}\rangle[\{1,2\}\ \mathbf{stit}]\mathrm{K}_1^{\circ\bullet\circ}p$$

[8] The formula $\langle Agt \setminus \{i\}\ \mathbf{stit}\rangle[Agt\ \mathbf{stit}]\neg\varphi$ expresses a *ceteris paribus* condition of the form "the choices of the agents in $Agt \setminus \{i\}$ being fixed, if i had chosen a different action then φ would have necessarily been false".

This means that at w there is no action such that if 1 *had chosen it then, after choosing it, agent* 1 *would have necessarily known that the paper was going to be accepted. The problem is that, before learning the choice of agent* 2*,* 1 *has a maximal uncertainty about it,* i.e.*,* 1 *does not know whether* 2 *is going to choose action S, action M or action L (see Figure 1b).*

Note that passive agentive responsibility implies passive causal responsibility, because of the truth axiom $\mathrm{K}_i^{\circ\bullet\circ}\varphi \rightarrow \varphi$ for *interim* knowledge. That is:

$$\models_{\text{E-GSTIT}} \mathrm{Resp}^{-}(i,\varphi) \rightarrow \mathrm{CHP}(i,\varphi) \tag{8}$$

Another interesting observation concerns the condition $\mathrm{K}_i^{\circ\bullet\circ}\varphi$ in the preceding definition of passive agentive responsibility. This condition prevents from considering an agent as agentively responsible for an outcome when the agent was not aware of the results of her choice. For example, let us suppose that Bill asks Ann to choose a restaurant for dinner. Ann decides to take him to her favourite restaurant. That night the main cook is ill and his assistant (who is not as good of a cook) is in charge of the cuisine. As a result, the food is not as good as usual and Bill is disappointed about the quality. Can we say that Ann is agentively responsible for Bill's bad experience? It does not seem to be the case, since Ann did not know that her choice would lead to that result (as she did not know that the main cook was ill).

Finally, let us point out that the reason why we do not define passive agentive responsibility as $\mathrm{K}_i^{\circ\bullet\circ}\mathrm{CHP}(i,\varphi)$ is that this definition would be too weak. Indeed, after making her choice, an agent i may know that she could have prevented φ from being true without knowing exactly which action could prevented φ from being true. This is related to the issue of *uniform choice* discussed by [26].

We conclude this section by considering some interesting validities about agentive responsibility. For $\tau \in \{+,-\}$ we have:

$$\models_{\text{E-GSTIT}} (\mathrm{Resp}^{\tau}(i,\varphi) \wedge \mathrm{Resp}^{\tau}(i,\psi)) \rightarrow \mathrm{Resp}^{\tau}(i,\varphi \wedge \psi) \tag{9}$$

According to this validity, if an agent is actively/passively agentive responsible for φ and she is actively/passively agentive responsible for ψ, then she is actively/passively agentive responsible for φ and ψ. For instance in order to prove the validity $\models_{\text{E-GSTIT}} (\mathrm{Resp}^{-}(i,\varphi) \wedge \mathrm{Resp}^{-}(i,\psi)) \rightarrow \mathrm{Resp}^{-}(i,\varphi \wedge \psi)$ we proceed as follows: $\mathrm{Resp}^{-}(i,\varphi)$ implies (1) $\mathrm{K}_i^{\circ\bullet\circ}\varphi$ while $\mathrm{Resp}^{-}(i,\psi)$ implies (2) $\mathrm{K}_i^{\circ\bullet\circ}\psi$. (1) and (2) together imply (3) $\mathrm{K}_i^{\circ\bullet\circ}(\varphi \wedge \psi)$. Moreover, $\mathrm{Resp}^{-}(i,\varphi)$ implies (4) $\langle Agt \setminus \{i\}\ \mathbf{stit}\rangle[Agt\ \mathbf{stit}]\mathrm{K}_i^{\circ\bullet\circ}(\neg\varphi \vee \neg\psi)$. (3) and (4) together are equivalent to $\mathrm{Resp}^{-}(i,\varphi \wedge \psi)$. Note that $\mathrm{Resp}^{-}(i,\varphi \wedge \psi) \rightarrow (\mathrm{Resp}^{-}(i,\varphi) \wedge \mathrm{Resp}^{-}(i,\psi))$ is not valid, because the fact that an agent could have prevented φ and ψ from being true *together* does not necessarily imply that both φ and ψ could have been avoided. For example, Bill could have prevented the state of affairs "Bill fails the exam *and* it is a sunny day" from being true, since he would have passed the exam if he had studied harder. Clearly, this does not imply that Bill could have prevented the state of affairs "it is a sunny day" from being true, as this state of affairs is completely independent from Bill's actual choice.

Moreover, $\mathrm{Resp}^{+}(i,\varphi \wedge \psi) \rightarrow (\mathrm{Resp}^{+}(i,\varphi) \wedge \mathrm{Resp}^{+}(i,\psi))$ is not valid because of the negative condition in the definition of the 'deliberative' STIT operator $[i\ \mathbf{dstit}]$. The following two validities highlight that an agent cannot be agentive responsible about tautologies or contradictions which seems highly intuitive. For $\tau \in \{+,-\}$ we have:

$$\models_{\text{E-GSTIT}} \neg\mathrm{Resp}^{\tau}(i,\top) \tag{10}$$

$$\models_{\text{E-GSTIT}} \neg\mathrm{Resp}^{\tau}(i,\bot) \tag{11}$$

4 Attribution emotions

In this section we use the concept of responsibility defined in Section 3 as departing point to formalize the cognitive structure of attribution emotions, *i.e.*, emotions such as guilt and pride which are based on the attribution of responsibility for a given outcome either to the self or to an external agent[9]. By 'cognitive structure' of an emotion, we mean the emotion's triggering conditions; that is, the agent's mental states that prompt the agent's emotional reaction and 'cause' the agent to feel the emotion.

We consider four different types of (cognitive structure of)[10] attribution emotions resulting from the combination of two different dimensions (see Table 2): internal *vs.* external attribution of responsibility [47]; and fulfillment *vs.* transgression of moral standards. Specifically, attribution emotions are either based (1) on the attribution of responsibility for the transgression of one's moral standards either to oneself (*i.e.*, guilt or self-directed moral disapproval) or to somebody else (*i.e.*, reproach or other-directed moral disapproval); or (2) on the attribution of responsibility for the fulfillment of one's moral standards either to oneself (*i.e.*, pride or self-directed moral approval) or to somebody else (*i.e.*, admiration or other-directed moral approval).

Table 2: The different kinds of attribution emotions

		Responsibility attribution	
		Self	Other
Moral standard	Fulfilment	Self-directed moral approval (Pride)	Other-directed moral approval (Admiration)
	Norm violation	Self-directed moral disapproval (Guilt)	Other-directed moral disapproval (Reproach)

Due to space restrictions we can not formalize the entire taxonomy of attribution emotions. Therefore, we only consider attribution emotions based on the violation of moral standards, namely guilt and reproach. Following current psychological theories of guilt (see, *e.g.*, [36, 21, 31]), we consider that the triggering condition of an agent i's guilt feeling is the agent i's knowledge that she is responsible for having behaved in a morally reprehensible way. Formally:

$$\mathrm{Guilt}(i) \stackrel{\text{def}}{=} \mathrm{K}_i^{\circ\circ\bullet}\mathrm{Resp}^{+}(i,viol_i) \vee \mathrm{K}_i^{\circ\circ\bullet}\mathrm{Resp}^{-}(i,viol_i)$$

The special atomic formula $viol_i$ is used to identify worlds which are considered sub-ideal by agent i; that is, those worlds that are in conflict with agent i's moral standards. Moreover, we consider that the triggering condition of an agent i's feeling of reproach towards some agent j is the agent i's knowledge that agent j is responsible for having behaved in a morally reprehensible way:

$$\mathrm{Reproach}(i,j) \stackrel{\text{def}}{=} \mathrm{K}_i^{\circ\circ\bullet}\mathrm{Resp}^{+}(j,viol_i) \vee \mathrm{K}_i^{\circ\circ\bullet}\mathrm{Resp}^{-}(j,viol_i)$$

[9]The term 'attribution emotion' is taken from Ortony *et al.* [37].

[10]Please note that for simplicity's sake and in order to improve readability, we will employ the expression "(attribution) emotion" for speaking of "cognitive structure of (attribution) emotion".

Note that in the preceding definitions we use the *ex post* knowledge operator $\mathrm{K}_i^{\circ\circ\bullet}$ because, in order to feel guilty, an agent i must reconsider the choice she has made retrospectively, that is, after the choices of the rest of the agents have been revealed. The reason why the interim knowledge operator $\mathrm{K}_i^{\circ\bullet\circ}$ is not sufficient is that agent i can evaluate the effective consequences of her choice, only after the choices of all agents have been revealed. By way of example, consider a typical coordination problem in which Bob and Mary, two university students, regularly meet in front of the university building at 12:30 a.m. in order to go to lunch together. One day Bob decides to go to lunch without waiting for Mary, because he is uncertain whether Mary will go to the usual meeting place. Indeed, Bob envisages the possibility that Mary decided to stay home to prepare for an exam. *Only after Bob discovers the actual choice of Mary*, namely that she came to the usual meeting place at 12:30 a.m., he does feel guilty for having let Mary wait for him. The use of the ex post knowledge operator $\mathrm{K}_i^{\circ\circ\bullet}$ in the definition of reproach can be justified in an analogous way.

As highlighted in the former definitions, we assume that the responsibility type does not play a role in the cognitive structures of guilt and reproach. In particular, we consider that a guilt feeling may be triggered by attribution of either active (*i.e.*, $\mathrm{K}_i^{\circ\circ\bullet}\mathrm{Resp}^{+}(i,viol_i)$) or passive (*i.e.*, $\mathrm{K}_i^{\circ\circ\bullet}\mathrm{Resp}^{-}(i,viol_i)$) responsibility to the self; whereas a reproach feeling can be triggered by attribution of either active (*i.e.*, $\mathrm{K}_i^{\circ\circ\bullet}\mathrm{Resp}^{+}(j,viol_i)$) or passive responsibility (*i.e.*, $\mathrm{K}_i^{\circ\circ\bullet}\mathrm{Resp}^{-}(j,viol_i)$) to another agent. Nevertheless, the responsibility type can play a role in the intensity of the emotion (a dimension that is however not studied further in this article). For instance, as shown by psychologists [29], people experience more regret over negative outcomes that stem from actions taken (regret due to action) than from equally negative outcomes resulting from actions foregone (regret due to inaction).

Before concluding, let us remark that an agent has positive introspection over her triggering condition of guilt as well as over her triggering condition of reproach:

$$\models_{\text{E-GSTIT}} \mathrm{Guilt}(i) \leftrightarrow \mathrm{K}_i^{\circ\circ\bullet}\mathrm{Guilt}(i) \tag{12}$$

$$\models_{\text{E-GSTIT}} \mathrm{Reproach}(i,j) \leftrightarrow \mathrm{K}_i^{\circ\circ\bullet}\mathrm{Reproach}(i,j) \tag{13}$$

Moreover, according to our definition, the triggering condition of guilt feeling is nothing but the triggering condition of self-reproach:

$$\models_{\text{E-GSTIT}} \mathrm{Guilt}(i) \leftrightarrow \mathrm{Reproach}(i,i) \tag{14}$$

5 From individual to collective

In sections 3 and 4 we have studied the relevance of the concept of responsibility attribution for emotion theory at the individual level. We now explore its collective dimension, and more specifically, the implications of collective responsibility for collective guilt attribution.

In order to make the qualitative leap from the individual to the group dimension of responsibility, there are two main questions we must address at the theoretical level:[11] *first*, the possibility for a group of individuals to cause something by choosing to perform a given joint action, *i.e.*, what we could denominate *'group choice condition'*; and *second*, a related account

[11] These questions are widely raised in the literature. See notably [18, p. 122], who states: "[t]he root worry over the idea of collective agentive responsibility may be a worry over the possibility over genuinely collective action and intention."

of group knowledge which allows us to distinguish a mere collective causal responsibility from a collective agentive responsibility *i.e.*, what we could denominate *'group knowledge condition'*. Indeed, in order to ascribe collective agentive responsibility to a given group of agents, it does not suffice that the agents cause something by performing a given joint action. The agents must also be *collectively* aware of the effects of their joint action.

The collective choice condition can be expressed in the logic E-GSTIT presented in Section 2. Indeed, in this logic one can express that (1) a group H 'deliberatively' sees to it that a state of affairs φ is true or that (2) a group H could have prevented φ from being true. In order to express sentences (1) and (2) we just need to generalize the 'deliberative' STIT operator $[i\ \mathbf{dstit}]$ and the construction $\mathrm{CHP}(i,\varphi)$ defined in Section 3 to groups as follows. For any $H \in 2^{Agt*}$:

$$[H\ \mathbf{dstit}]\varphi \stackrel{\text{def}}{=} [H\ \mathbf{stit}]\varphi \wedge \langle \emptyset\ \mathbf{stit}\rangle \neg\varphi$$

$$\mathrm{CHP}(H,\varphi) \stackrel{\text{def}}{=} \varphi \wedge \langle Agt \setminus H\ \mathbf{stit}\rangle [Agt\ \mathbf{stit}]\neg\varphi$$

$[H\ \mathbf{dstit}]\varphi$ and $\mathrm{CHP}(H,\varphi)$ capture respectively group H's active causal responsibility for making φ true and group H's passive causal responsibility for letting φ be true.

On the contrary, the preceding group knowledge condition cannot be expressed in the logic E-GSTIT of Section 2. What the logic E-GSTIT misses is a modal operator of group knowledge, which we present hereupon.

In the present work we assimilate the notion of group knowledge to the concept of common knowledge which is widely used both in computer science [14] and in economics [4]. The distinguishing feature of the notion of common knowledge is its reductionist view of group knowledge, in the sense that common knowledge implies shared knowledge: when there is a common knowledge in group J that φ is true, then each agent in J individually knows that φ is true. Alternative "non-reductionist" notions of group knowledge have been studied both in the philosophical literature on group attitudes [17, 41, 40] and in the logical literature [34].[12] One of these "non-reductionist" notions is collective acceptance [40], according to which a group of individuals may *collectively* accept that φ is true even though some of its members do not believe so *individually*. For example, the members of a Parliament might collectively accept (*qua* members of the Parliament) that launching a military action against another country is legitimate because the Parliament's majority has decided so, even though some of them - who voted against the military intervention - individually believe the contrary.

We introduce a new logic, E-GSTIT$^+$, which extends E-GSTIT with three different types of common knowledge modal operators, corresponding to the three different types of epistemic operators presented in Section 2: *ex ante* common knowledge, *interim* common knowledge and *ex post* common knowledge. Specifically, the language $\mathcal{L}_{\text{E-GSTIT}^+}(Atm, Agt)$ of the logic E-GSTIT$^+$ is the set of formulae defined by the following BNF:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid [J\ \mathbf{stit}]\varphi \mid \mathrm{K}_i^{\bullet\circ\circ}\varphi \mid \mathrm{K}_i^{\circ\bullet\circ}\varphi \mid$$
$$\mathrm{K}_i^{\circ\circ\bullet}\varphi \mid \mathrm{CK}_H^{\bullet\circ\circ}\varphi \mid \mathrm{CK}_H^{\circ\bullet\circ}\varphi \mid \mathrm{CK}_H^{\circ\circ\bullet}\varphi$$

where p ranges over Atm, J ranges over 2^{Agt}, H ranges over 2^{Agt*} and i ranges over Agt.

The truth conditions of E-GSTIT$^+$ formulae are the ones given in Definition 5, plus the following three additional conditions for *ex-ante*, *interim* and *ex-post* common knowledge.

[12] Although extremely important, this debate between reductionist vs. non-reductionist views of group knowledge escapes the scope of this article. See [33] for a synthetic overview of this debate.

Definition 6 (Truth conditions (cont.)) *Let M be an ESCN and w a world in M:*

$$\begin{array}{lll} M,w \models \mathrm{CK}_H^{\bullet\circ\circ}\varphi & \iff & \forall v \in W \textit{ such that } w(\mathcal{E}_H^{\bullet\circ\circ})^+ v : M,v \models \varphi \\ M,w \models \mathrm{CK}_H^{\circ\bullet\circ}\varphi & \iff & \forall v \in W \textit{ such that } w(\mathcal{E}_H^{\circ\bullet\circ})^+ v : M,v \models \varphi \\ M,w \models \mathrm{CK}_H^{\circ\circ\bullet}\varphi & \iff & \forall v \in W \textit{ such that } w(\mathcal{E}_H^{\circ\circ\bullet})^+ v : M,v \models \varphi \end{array}$$

where for $x \in \{\bullet\circ\circ, \circ\bullet\circ, \circ\circ\bullet\}$, $\mathcal{E}_H^x = \bigcup_{i\in H} \mathcal{E}_i^x$ and $(\mathcal{E}_H^x)^+$ is the transitive closure of $\mathcal{E}_H^x$.

For any formula φ of the language $\mathcal{L}_{\text{E-GSTIT}^+}(Atm, Agt)$, we write $\models_{\text{E-GSTIT}^+} \varphi$ if φ is E-GSTIT$^+$ *valid*, that is, if φ is true in all ESCNs (*i.e.*, for all ESCNs M and for all worlds w in M, we have $M,w \models \varphi$). We say that φ is E-GSTIT$^+$ *satisfiable* if $\neg\varphi$ is not E-GSTIT$^+$ valid.

In logic E-GSTIT$^+$ we can finally draw the distinction between the notion of collective causal responsibility defined above and the notion of collective agentive responsibility. This distinction parallels the one made in Section 3 between individual causal responsibility and individual agentive responsibility.

For the active form of collective agentive responsibility we define:

$$\mathrm{Resp}^+(H,\varphi) \stackrel{\text{def}}{=} \mathrm{CK}_H^{\circ\bullet\circ}[H\ \mathbf{dstit}]\varphi$$

This means that the agents in group H are together agentively responsible (in the active sense) for bringing about φ (denoted by $\mathrm{Resp}^+(H,\varphi)$) if and only if, there is common knowledge in group H that the group makes φ true and that φ is not something inevitable. In other words, group H *knowingly* sees to it that φ.

We employ the *interim* common knowledge operator $\mathrm{CK}_H^{\circ\bullet\circ}$ instead of the *ex ante* common knowledge operator $\mathrm{CK}_H^{\bullet\circ\circ}$ or the *ex post* common knowledge operator $\mathrm{CK}_H^{\circ\circ\bullet}$ because, similarly to the case of individual agentive responsibility, in order to be agentively responsible for something, the agents in a group have to be mutually aware of the potential effects of their joint action after having made their choice but before knowing the choices of the agents outside the group.

For the passive form of collective agentive responsibility we define:

$$\mathrm{Resp}^-(H,\varphi) \stackrel{\text{def}}{=} \mathrm{CK}_H^{\circ\bullet\circ}\varphi \wedge \langle Agt \setminus H\ \mathbf{stit}\rangle[Agt\ \mathbf{stit}]\mathrm{CK}_H^{\circ\bullet\circ}\neg\varphi$$

This means that the agents in group H are together responsible (in the passive sense) for letting φ be true if and only if: (1) they have common knowledge that their current joint action leads to φ (*i.e.*, $\mathrm{CK}_H^{\circ\bullet\circ}\varphi$) and (2) given what the agents outside H have decided to do, if the agents in H had chosen a different joint action, then they would have necessarily acquired common knowledge of the contrary (*i.e.*, $\langle Agt\setminus H\ \mathbf{stit}\rangle[Agt\ \mathbf{stit}]\mathrm{CK}_H^{\circ\bullet\circ}\neg\varphi$). Note that the passive form of collective agentive responsibility implies the passive form of collective causal responsibility, because of the truth axiom $\mathrm{CK}_H^{\circ\bullet\circ}\varphi \to \varphi$ for *interim* common knowledge. That is:

$$\models_{\text{E-GSTIT}^+} \mathrm{Resp}^-(H,\varphi) \to \mathrm{CHP}(H,\varphi) \tag{15}$$

Another interesting observation concerns the relationship between collective agentive responsibility and individual agentive responsibility. According to our definitions, the former does not necessarily imply the latter, in the sense that a group of agents may be collectively responsible for φ, either in the active or the passive sense, even though some of its

members are not. Indeed, the following E-GSTIT$^+$ formulae are satisfiable with $i \in H$: $\mathrm{Resp}^+(H,\varphi) \wedge \neg\mathrm{Resp}^+(i,\varphi)$ and $\mathrm{Resp}^-(H,\varphi) \wedge \neg\mathrm{Resp}^-(i,\varphi)$. The reason why these formulae are satisfiable lays in the group choice condition. In particular, in the logic E-GSTIT the fact that a group H causes φ to be true (*i.e.*, $[H\ \mathbf{dstit}]\varphi$) does not necessarily imply that each agent in H causes φ to be true (*i.e.*, $[i\ \mathbf{dstit}]\varphi$ with $i \in H$) , and the fact that a group H could have prevented φ from being true (*i.e.*, $\mathrm{CHP}(H,\varphi)$) does not necessarily imply that each agent in H could have prevented φ from being true (*i.e.*, $\mathrm{CHP}(i,\varphi)$ with $i \in H$).

Once we have made the leap from the individual to the collective dimension of responsibility attribution, we can make the link with the concept of collective guilt. Following Gilbert's philosophical account of collective guilt [18], we say that the agents in a group H feel *together* guilty (denoted by $\mathrm{Guilt}(H)$) if and only if the agents in H have common knowledge that they are together responsible for having behaved in a morally reprehensible way:

$$\mathrm{Guilt}(H) \stackrel{\text{def}}{=} \mathrm{CK}_H^{\circ\circ\bullet}\mathrm{Resp}^+(H,viol_H) \vee \mathrm{CK}_H^{\circ\circ\bullet}\mathrm{Resp}^-(H,viol_H)$$

where $viol_H \stackrel{\text{def}}{=} \bigwedge_{i\in H} viol_i$. The construction $viol_H$ is used to identify worlds which are considered subideal by all agents in the group H; that is, those worlds that are in conflict with the moral standards of all agents in H.

As the following E-GSTIT$^+$ validity highlights, a group has positive introspection over its collective guilt, in the sense that the agents in a group feel *together* guilty, if and only if they have common knowledge of the following:

$$\models_{\text{E-GSTIT}^+} \mathrm{Guilt}(H) \leftrightarrow \mathrm{CK}_H^{\circ\circ\bullet}\mathrm{Guilt}(H) \tag{16}$$

Gilbert makes the distinction between the preceding type of collective guilt and what she calls *membership guilt*, where an agent feels guilty because of her belonging to a group which is responsible for having behaved in a morally reprehensible way. This notion of membership guilt is expressed by the E-GSTIT$^+$ formula:

$$\mathrm{K}_i^{\circ\circ\bullet}\mathrm{Resp}^+(H,viol_i) \vee \mathrm{K}_i^{\circ\circ\bullet}\mathrm{Resp}^-(H,viol_i) \qquad \text{for } i \in H$$

In the following section we deal with more technical issues by studying the computational complexity of the logic E-GSTIT$^+$.

6 Complexity issues

There are a number of questions that are worth mentioning regarding decidability of the logics presented in Sections 2 and 5. *Firstly,* we must note that the satisfiability problem of the logic GSTIT interpreted over Kripke group STIT models (see Definition 2) is undecidable if there are more than 2 agents. This result has been proved by [25], where it is shown that the logic GSTIT with n agents interpreted over Kripke group STIT models is just the product logic S5^n whose satisfiability problem has been proved to be undecidable when $n \geq 3$ [15]. *Secondly,* from Theorem 1 given in Section 2.2, one can prove the following corollary about the connection between the notion of GSTIT satisfiability with respect to the class of STIT models with choice names (SCNs) and the notion of GSTIT satisfiability with respect to the class of Kripke group STIT models.

Corollary 1 *Let φ be a formula of the language $\mathcal{L}_{GSTIT}(Atm, Agt)$ and let $card(Act) = n$ with $n \in \mathbb{N} \cup \{\infty\}$. Then, there exists a SCN M and a world w in M such that $M, w \models \varphi$ if and only if there exists a Kripke group STIT model M' whose number of choices is bounded by n and a world w' in M' such that $M', w' \models \varphi$.*

It follows that, if $card(Act) = \infty$, then the notion of GSTIT satisfiability with respect to the class of SCNs and the notion of GSTIT satisfiability with respect to the class of Kripke group STIT models are equivalent. Consequently, if $card(Act) = \infty$ and $card(Agt) \geq 3$, then the problem of testing satisfiability of a GSTIT formula with respect to the class of SCNs is undecidable too.

Therefore, since the logic E-GSTIT is nothing but an extension of the logic GSTIT, from Corollary 1 it follows that, if the set of action terms Act is infinite and there are more than 2 agents in Agt, then the satisfiability problems of the logics E-GSTIT and E-GSTIT$^+$ are also undecidable. This is the reason why we are interested in studying the decidability of E-GSTIT and E-GSTIT$^+$ under the assumption that the set of action terms Act is finite.

In what follows we focus on the complexity of the satisfiability problem of E-GSTIT$^+$ under the assumption that Act is finite.

We are going to prove that, if Act is finite, the logic E-GSTIT$^+$ can be embedded into the decidable logic S5-PDL; *i.e.*, the variant of propositional dynamic logic PDL [22] in which atomic programs are interpreted by means of equivalence relations.

Let $Atm^+ = Atm \cup \{\mathrm{ch}_{i:a} | i \in Agt \text{ and } a \in Act\}$. The special atomic propositions $\mathrm{ch}_{i:a}$ are used to describe the action chosen by an agent at a given world. In particular, $\mathrm{ch}_{i:a}$ has to be read "the agent i chooses the action a". For all $H \in 2^{Agt*}$ and for all $\alpha_H \in Act_H$ we define:

$$\mathrm{ch}_{\alpha_H} \stackrel{\text{def}}{=} \bigwedge_{i \in H} \mathrm{ch}_{i:\alpha_H(i)}$$

where ch_{α_H} has to be read "the agents in the coalition H choose the joint action α_H".

The language $\mathcal{L}_{\text{S5-PDL}}(Atm^+, Agt)$ of S5-PDL is defined as follows:

$$\pi ::= \equiv | \sim_i | \ \pi_1;\pi_2 \mid \pi_1 \cup \pi_2 \mid \pi^* \mid ?\varphi$$
$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid [\pi]\varphi$$

where p ranges over Atm^+ and i ranges over Agt. The atomic program $\equiv$ is used to simulate the STIT operator $[\emptyset\ \mathbf{stit}]$ for the empty coalition, whereas the atomic program $\sim_i$ is used to simulate the *ex ante* knowledge operator $\mathrm{K}_i^{\bullet\circ\circ}$ for agent i. The dual of the operator $[\pi]$ is defined in the standard way as follows: $\langle\pi\rangle\varphi \stackrel{\text{def}}{=} \neg[\pi]\neg\varphi$.

S5-PDL models are tuples $M = \langle W, \mathcal{R}_{\equiv}, \{\mathcal{R}_{\sim_i}\}_{i \in Agt}, \mathcal{I}\rangle$ where:

- W is a set of worlds,
- $\mathcal{R}_{\equiv}$ and all $\mathcal{R}_{\sim_i}$ are equivalence relations on W,
- $\mathcal{I} : Atm^+ \longrightarrow 2^W$ is a valuation function.

Binary relations for complex programs are defined in the standard way as follows:

$$\mathcal{R}_{\pi_1;\pi_2} = \mathcal{R}_{\pi_1};\mathcal{R}_{\pi_2}$$
$$\mathcal{R}_{\pi_1 \cup \pi_2} = \mathcal{R}_{\pi_1} \cup \mathcal{R}_{\pi_2}$$
$$\mathcal{R}_{\pi^*} = (\mathcal{R}_{\pi})^*$$
$$\mathcal{R}_{?\varphi} = \{(w, w) | w \in W \text{ and } M, w \models \varphi\}$$

Truth conditions of S5-PDL formulae are the standard ones for Boolean operators plus the following one for the operator $[\pi]$:

$$M, w \models [\pi]\varphi \iff \forall v \in \mathcal{R}_\pi(w) : M, v \models \varphi$$

For any formula φ of the language $\mathcal{L}_{\text{S5-PDL}}(Atm^+, Agt)$, we write $\models_{\text{S5-PDL}} \varphi$ if φ is S5-PDL *valid*, that is, if φ is true in all S5-PDL models (*i.e.*, for all S5-PDL models M and for all worlds w in M, we have $M, w \models \varphi$). We say that φ is S5-PDL *satisfiable* if $\neg\varphi$ is not S5-PDL valid. Moreover, we shall say that φ is a global logical consequence in S5-PDL of a finite set of global axioms $\Gamma = \{\chi_1, \ldots, \chi_n\}$, denoted by $\Gamma \models_{\text{S5-PDL}} \varphi$, if and only if for every S5-PDL model M, if Γ is true in M (*i.e.*, for every world w in M, we have $M, w \models \chi_1 \wedge \ldots \wedge \chi_n$) then φ is true in M too (*i.e.*, for every world w in M, we have $M, w \models \varphi$).

Proposition 1 *The satisfiability problem of S5-PDL is in ExpTime.*

PROOF. The logic S5-PDL is polynomially embeddable into PDL extended with converse, by simulating S5 programs $\equiv$ and $\sim_i$ with composite programs $(x \cup -x)^*$ and $(x_i \cup -x_i)^*$ where x and x_i are arbitrary atomic programs interpreted by means of binary relations $\mathcal{R}_x$ and $\mathcal{R}_{x_i}$, $-x$ is the converse of x and $-x_i$ is the converse of x_i . (Note indeed that relations $\mathcal{R}_{(x\cup -x)^*}$ and $\mathcal{R}_{(x_i\cup -x_i)^*}$ are equivalence relations.) The satisfiability problem of PDL with converse has been proved to be ExpTime-complete [44]. It follows that the satisfiability problem of S5-PDL is in ExpTime. ■

We define the following translation from the E-GSTIT$^+$ language $\mathcal{L}_{\text{E-GSTIT}^+}(Atm, Agt)$ to $\mathcal{L}_{S5-\text{PDL}}(Atm^+, Agt)$ for $p \in Atm$ and $H \in 2^{Agt*}$:

$$\begin{aligned}
tr(p) &= p \\
tr(\neg\varphi) &= \neg tr(\varphi) \\
tr(\varphi \wedge \psi) &= tr(\varphi) \wedge tr(\psi) \\
tr([\emptyset\ \mathbf{stit}]\varphi) &= [\equiv] tr(\varphi) \\
tr([H\ \mathbf{stit}]\varphi) &= \Big[\bigcup_{\alpha_H \in Act_H} (?\mathrm{ch}_{\alpha_H}; \equiv ; ?\mathrm{ch}_{\alpha_H}) \Big] tr(\varphi) \\
tr(\mathrm{K}_i^{\bullet\circ\circ}\varphi) &= [\sim_i] tr(\varphi) \\
tr(\mathrm{K}_i^{\circ\bullet\circ}\varphi) &= \Big[\bigcup_{a \in Act} (?\mathrm{ch}_{i:a}; \sim_i ; ?\mathrm{ch}_{i:a}) \Big] tr(\varphi) \\
tr(\mathrm{K}_i^{\circ\circ\bullet}\varphi) &= \Big[\bigcup_{\alpha_{Agt} \in Act_{Agt}} (?\mathrm{ch}_{\alpha_{Agt}}; \sim_i ; ?\mathrm{ch}_{\alpha_{Agt}}) \Big] tr(\varphi) \\
tr(\mathrm{CK}_H^{\bullet\circ\circ}\varphi) &= \Big[(\bigcup_{i \in H} \sim_i)^* \Big] tr(\varphi) \\
tr(\mathrm{CK}_H^{\circ\bullet\circ}\varphi) &= \Big[(\bigcup_{i \in H, a \in Act} (?\mathrm{ch}_{i:a}; \sim_i ; ?\mathrm{ch}_{i:a}))^* \Big] tr(\varphi) \\
tr(\mathrm{CK}_H^{\circ\circ\bullet}\varphi) &= \Big[(\bigcup_{i \in H, \alpha_{Agt} \in Act_{Agt}} (?\mathrm{ch}_{\alpha_{Agt}}; \sim_i ; ?\mathrm{ch}_{\alpha_{Agt}}))^* \Big] tr(\varphi)
\end{aligned}$$

As the following lemma highlights, the validity problem in E-GSTIT$^+$ can be reduced to the problem of (global) logical consequence in S5-PDL.

Lemma 1 *A E-GSTIT$^+$ formula φ is E-GSTIT$^+$ valid,* i.e., $\models_{\textit{E-GSTIT}^+} \varphi$*, if and only if $tr(\varphi)$ is a logical consequence of the set of global axioms $\Gamma_{Agt,Act}$ in S5-PDL,* i.e., $\Gamma_{Agt,Act} \models_{\textit{S5-PDL}} tr(\varphi)$*, where:*

$$\begin{aligned}\Gamma_{Agt,Act} \;=\; & \Big\{ \bigvee_{a\in Act} \mathrm{ch}_{i:a} | i \in Agt \Big\} \cup \\ & \big\{ \mathrm{ch}_{i:a} \rightarrow \neg \mathrm{ch}_{i:b} | i \in Agt, a, b \in Act \textit{ and } a \neq b \big\} \cup \\ & \big\{ (\langle\equiv\rangle \mathrm{ch}_{1:\alpha_{Agt}(1)} \wedge \ldots \wedge \langle\equiv\rangle \mathrm{ch}_{n:\alpha_{Agt}(n)}) \rightarrow \langle\equiv\rangle \mathrm{ch}_{\alpha_{Agt}} | \alpha_{Agt} \in Act_{Agt} \big\}\end{aligned}$$

PROOF.(Sketch) ($\Rightarrow$) Take an arbitrary ESCN $M = \langle W, \{\mathcal{A}_i\}_{i\in Agt}, \{\mathcal{R}_J\}_{J\subseteq Agt}, \{\mathcal{E}_i^{\bullet\circ\circ}, \mathcal{E}_i^{\circ\bullet\circ}, \mathcal{E}_i^{\circ\circ\bullet}\}_{i\in Agt}, \mathcal{V}\rangle$ which satisfies formula φ. We build a corresponding S5-PDL model $M' = \langle W, \mathcal{R}_\equiv, \{\mathcal{R}_{\sim_i}\}_{i\in Agt}, \mathcal{I}\rangle$ such that:

- $\mathcal{R}_\equiv = \mathcal{R}_\emptyset$;
- for all $i \in AGT$, $\mathcal{R}_{\sim_i} = \mathcal{E}_i^{\bullet\circ\circ}$;
- for all $p \in Atm$, $\mathcal{I}(p) = \mathcal{V}(p)$;
- for all $i \in Agt$ and for all $a \in Act$, $\mathcal{I}(\mathrm{ch}_{i:a}) = \{w \in W | \mathcal{A}_i(w) = a\}$.

By induction on the structure of φ, it is easy to check that $tr(\varphi)$ is satisfied by M' and that for every $\chi \in \Gamma$, $M', w \models \chi$ for all $w \in W$.

($\Leftarrow$) Take an arbitrary S5-PDL model $M = \langle W, \mathcal{R}_\equiv, \{\mathcal{R}_{\sim_i}\}_{i\in Agt}, \mathcal{I}\rangle$ which satisfies $tr(\varphi)$ such that that for all $\chi \in \Gamma$ and for all $w \in W$, $M, w \models \chi$. We build a corresponding ESCN $M' = \langle W, \{\mathcal{A}_i\}_{i\in Agt}, \{\mathcal{R}_J\}_{J\subseteq Agt}, \{\mathcal{E}_i^{\bullet\circ\circ}, \mathcal{E}_i^{\circ\bullet\circ}, \mathcal{E}_i^{\circ\circ\bullet}\}_{i\in Agt}, \mathcal{V}\rangle$ such that:

- for all $i \in Agt$ and for all $a \in Act$, $\mathcal{A}_i(w) = a$ if and only if $w \in \mathcal{I}(\mathrm{ch}_{i:a})$;
- $\mathcal{R}_\emptyset = \mathcal{R}_\equiv$;
- for all $H \in 2^{Agt*}$, $\mathcal{R}_H = \{(w, v) | w\mathcal{R}_\emptyset v$ and $\mathcal{A}_H(w) = \mathcal{A}_H(v)\}$;
- for all $i \in Agt$, $\mathcal{E}_i^{\bullet\circ\circ} = \mathcal{R}_{\sim_i}$;
- for all $i \in Agt$, $\mathcal{E}_i^{\circ\bullet\circ} = \{(w, v) | w\mathcal{E}_i^{\bullet\circ\circ} v$ and $\mathcal{A}_i(w) = \mathcal{A}_i(v)\}$;
- for all $i \in AGT$, $\mathcal{E}_i^{\circ\circ\bullet} = \{(w, v) | w\mathcal{E}_i^{\bullet\circ\circ} v$ and $\mathcal{A}_{Agt}(w) = \mathcal{A}_{Agt}(v)\}$;
- for all $p \in Atm$, $\mathcal{V}(p) = \mathcal{I}(p)$.

Again, by induction on the structure of φ, it is easy to check that φ is satisfied by M'. ■

From Proposition 1 and Lemma 1 we obtain the following complexity result for E-GSTIT$^+$ under the assumption that the set of action terms Act is finite.

Theorem 2 *Let Act be a finite set. Then, the satisfiability problem of E-GSTIT$^+$ is decidable in exponential time.*

PROOF. The logic E-GSTIT$^+$ is an extension of the logic of common knowledge which has been proved to be ExpTime-complete [14]. This provides an argument for the ExpTime-hardness of the satisfiability problem of E-GSTIT$^+$.

It is a routine task to verify that the problem of global logical consequence in S5-PDL with a finite number of global axioms is reducible to the problem of validity in S5-PDL. In

particular, if $\Gamma = \{\chi_1, \ldots, \chi_m\}$ we have $\Gamma \models_{\text{S5-PDL}} \varphi$ if and only if $\models_{\text{S5-PDL}} [\mathbf{any}^*](\chi_1 \wedge \ldots \chi_m) \rightarrow \varphi$ where $\mathbf{any}$ is the special program defined as $\mathbf{any} \stackrel{\text{def}}{=} (\bigcup_{i \in Agt} \sim_i \cup \equiv)$. Hence, by Lemma 1 and by the fact that $\Gamma_{Agt,Act}$ is finite (because Act and Agt are finite), it follows that a E-GSTIT$^+$ formula φ is E-GSTIT$^+$ valid if and only if $\models_{\text{S5-PDL}} [\mathbf{any}^*](\bigwedge_{\chi \in \Gamma_{Agt,Act}} \chi) \rightarrow tr(\varphi)$. Consequently, given the fact that tr is a polynomial reduction of E-GSTIT$^+$ to S5-PDL and given also the fact that the satisfiability problem of S5-PDL is in ExpTime (Proposition 1), it follows that the satisfiability problem of E-GSTIT$^+$ is also in ExpTime. ■

Given that E-GSTIT$^+$ is an extension of E-GSTIT, Theorem 2 also provides a decidability result for the logic E-GSTIT. We conjecture that the satisfiability problem of the logic E-GSTIT with Act finite is PSPACE-complete. Unfortunately, we shall postpone an in-depth analysis of this interesting issue to future work.

7 Related Work

There have been other attempts to combine the logic STIT with epistemic modalities (see, *e.g.*, [10, 8, 26]) in the past. However, to our knowledge, the present work is the first attempt to combine group STIT with (1) three types of epistemic operators for *ex ante*, *interim* and *ex post* knowledge and (2) their corresponding operators of *ex ante*, *interim* and *ex post* common knowledge, providing (3) an analysis of the computational complexity of this extension of STIT. As emphasized in Section 2, the distinction between *ex ante*, *interim* and *ex post* knowledge is commonly used in the theory of normal form games [3]. Since basic group STIT without knowledge and common knowledge is already undecidable, in the present work we have focused on an epistemic extension of group STIT under the assumption that agents' available choices are finite. It is worth noting that a similar idea of formalizing *ex ante*, *interim* and *ex post* knowledge in STIT is sketched by [7]. However, his analysis is more restrictive than ours since he assumes perfect information, whereas our logic E-GSTIT allows for uncertainty over the states of the world. Furthermore, [7] does not consider common knowledge or complexity issues for this STIT extension.

The additional original contribution of the present work is the alternative semantics for group STIT in terms of STIT models with choice names (SCNs). As underlined in Section 2, SCNs provide an explicit action-based version of the semantics for group STIT. An interesting aspect of SCNs is that it provides a representation of social interaction that is very close to the game-theoretic representation of social interaction (in which actions and strategies of agents and coalitions are explicit).

On the conceptual side, the formalization we propose in this article represents the first logical analysis, based on STIT logic, of attribution emotions (and more concretely guilt) that deals with the responsibility aspect of such phenomena.

Our formalization of active and passive responsibility also constitutes a novelty, as earlier works mainly focus on the active aspect of responsibility. For instance, we can find a similar notion of moral responsibility in [32, p. 110] where, in order "[...] to be blamed for φ, the agent must have the intention to attain φ, or at least, intentionally let it happen." However, the latter do not include in their definition the passive aspect of responsibility. Furthermore, Lima et al.'s [32, p. 101] formalization of responsibility also introduces the notion of agent's knowledge,

defining for each agent an equivalence relation between possible worlds. Still, they do not specify the temporal nature of such knowledge with respect to the agent's choice of action, that is, the notion of knowledge introduced allows them to express agent's incomplete knowledge about the situation, but in a general way. Our E-GSTIT introduces three different concepts and corresponding modal operators of knowledge: $\mathrm{K}_i^{\bullet\circ\circ}$, $\mathrm{K}_i^{\circ\bullet\circ}$ and $\mathrm{K}_i^{\circ\circ\bullet}\varphi$, which add substantial profundity to the analysis and allows to capture subtleties that a general knowledge operator may not.

Other works, such as [13], present a formalization of responsibility and other organization-related concepts, where organizations are viewed as 'instances of normative systems' formed of agents' interaction patterns obeying the rules stated by a normative system. Organizations, and therefore responsibilities, are analyzed from the point of view of the rules to which the organized group of agents is subjected. The latter authors do include the passive aspect of responsibility, in the sense that "the agent could have prevented φ from happening", and chained responsibility for others' actions ('responsibility for other's agency' and 'responsibility with possibility of delegation'). However, they do not include the collective aspect of responsibility. In addition, Cholvy et al. [13] do not make the distinction between the three types of knowledge (*ex ante*, *interim* and *ex post*) that is crucial for the analysis of responsibility attribution.

In this line of research, we can also refer to [20]. While Grossi et al. are able to handle in quite fine-grained details the notion of plan having at disposals sets of atomic actions, they do not capture the collective aspect of responsibility, but rather indirect responsibility in the sense of delegation, or influence, likewise to [13]. Agents' activities are captured via "bringing-it-about" modal constructs and organizational activities (in particular delegation) are described via modal operators modeling forms of indirect action or influence ("bringing-it-about indirectly that").

Furthermore, most of the existing formalizations do not contemplate the role of responsibility attribution in the theory of attribution emotions. For instance, the formal representation presented in [1] does not include guilt, and [39] does not present a model of attribution emotions based on a systematic logical analysis of causal and moral responsibility.

Therefore, although there have certainly been other attempts to investigate the concept of responsibility in multi-agent systems with the help of logical methods, up to now there is no comprehensive logical study of the following two aspects of responsibility attribution: 1) *the individual and the collective dimensions of responsibility attribution* (*i.e.*, an agent ascribing responsibility for an action to herself or to another agent *vs.* a group of agents ascribing responsibility for an action to itself or to another group of agents); 2) *the active and passive aspects of responsibility*; and 3) *the role of responsibility attribution in the theory of attribution emotions*. Hence, the aim of this paper is to fill this gap in the logical literature on responsibility.

8 Conclusion

We have provided a formal analysis of the concept of responsibility attribution with the help of an extension of the logic STIT with modal operators for individual and common knowledge. Furthermore, we have studied decidability of the satisfiability problems of our logics E-GSTIT and E-GSTIT$^+$ under the assumption that agents' choices are finite. More specifically, we have proved that the complexity of the satisfiability problem of E-GSTIT$^+$ is ExpTime-complete.

The general insight gained from our paper is that a fine-grained and comprehensive analysis

of the concept of agentive responsibility (both along the dimensions active vs. passive and individual vs. collective) and of the related concept of attribution emotion requires a distinction between the three types of *ex ante*, *interim* and *ex post* knowledge. For example, as shown in Section 3, the operator of *interim* knowledge is required to formally characterize the concept of active agentive responsibility. Indeed, in order to define active agentive responsibility, we need to represent the agent's epistemic state after she has made her choice. The main innovation of our extension of STIT lies therefore in the possibility of distinguishing these three forms of knowledge.

Directions of future research are manifold. On the technical side, we plan to study the complexity of the satisfiability problem of the less expressive logic E-GSTIT. Another issue for future research is the problem of finding sound and complete axiomatizations for both E-GSTIT and E-GSTIT$^+$.

On the conceptual side, we are aware that this paper represents only the first step towards a more general logical account of the individual and collective aspects of responsibility attribution. Henceforth, we plan to enrich our analysis of collective emotions by considering emotion types different from collective guilt, such as collective pride and collective reproach. We also plan to investigate the logical relationship between guilt and shame. This topic has been already studied in the past by [42] in a framework based on dynamic logic. We consider that our approach based on STIT logic can provide interesting insights on the role responsibility attribution plays in the cognitive structure of shame.

Finally, we defer to future research an extension of the basic framework E-GSTIT with a weaker or less demanding notion of group knowledge than that of common knowledge used in the present article, such as for example distributed knowledge or the non-reductionist notion of collective acceptance studied by [34].

9 Acknowledgements

This research has been supported by the French ANR project EmoTES "Emotions in strategic interaction: theory, experiments, logical and computational studies", contract No. 11-EMCO-004-01.

References

[1] C. Adam. *The emotions: from psychological theories to logical formalisation and implementation in a BDI agent*. PhD thesis, Toulouse Research Institute in Computer Science, Universite Paul Sabatier, 2007.

[2] Aristotle. *Nicomachean Ethics*, volume 3, chapter 5: Moral virtue. CreateSpace Independent Publishing Platform, 2011.

[3] R. Aumann and J. Dreze. Rational expectations in games. *American Economic Review*, 98(1):72–86, 2008.

[4] R. J. Aumann. Backward induction and common knowledge of rationality. *Games and Economic Behavior*, 8:69, 1995.

[5] P. Balbiani, A. Herzig, and N. Troquard. Alternative axiomatics and complexity of deliberative stit theories. *Journal of Philosophical Logic*, 37(4):387–406, 2008.

[6] N. Belnap, M. Perloff, and M. Xu. *Facing the future: agents and choices in our indeterminist world*. Oxford University Press, USA, 2001.

[7] M. M. Bentzen. *Stit, Iit, and Deontic Logic for Action Types*. PhD thesis, Roskylde University, 2010.

[8] J. Broersen. A complete STIT logic for knowledge and action, and some of its applications. In *Proceedings of the 6th International Workshop on Declarative Agent Languages and Technologies (DALT 2008)*, volume 5397 of *LNCS*, pages 47–59. Springer-Verlag, 2008.

[9] J. Broersen. A logical analysis of the interaction between 'obligation-to-do' and 'knowingly doing'. In R. van der Meyden and L. van der Torre, editors, *Proceedings of the 9th International Conference on Deontic Logic in Computer Science (DEON 2008)*, volume 5076 of *Lecture Notes in Computer Science*, pages 140–154. Springer, 2008.

[10] J. Broersen. Deontic epistemic stit logic distinguishing modes of mens rea. *Journal Applied Logic*, 9(2), 2011.

[11] B. J. Chellas. Time and modality in the logic of agency. *Studia Logica*, 51:485–517, 1992.

[12] R. Chisholm. Freedom and action. In K. Lehrer, editor, *Freedom and Determinism*, pages 28–44. Random House, Oxford, 1966.

[13] L. Cholvy, F. Cuppens, and C. Saurel. Towards a logical formalization of responsibility. In *Proceedings of ICAIL 1997*, pages 233–242, 1997.

[14] R. Fagin, J. Y. Halpern, Y. Moses, and M. Vardi. *Reasoning about Knowledge*. MIT Press, Cambridge, 1995.

[15] D. M. Gabbay, A. Kurucz, F. Wolter, and M. Zakharyaschev. *Many-dimensional modal logics: theory and applications*. Elsevier, 2003.

[16] J. Gerbrandy. Logics of propositional control. In *Proceedings of AAMAS'06*, pages 193–200. ACM, 2006.

[17] M. Gilbert. *On Social Facts*. Routledge, London and New York, 1989.

[18] M. Gilbert. Collective guilt and collective feelings. *The Journal of Ethics*, 6:115–143, 2002.

[19] D. Gotterbarn. Informatics and professional responsibility. *Science and Engineering Ethics*, 7:221–230, 2001.

[20] D. Grossi, L. M. M. Royakkers, and F. Dignum. Organizational structure and responsibility. *Artificial Intelligence and Law*, 15(3):223–249, 2007.

[21] J. Haidt. The moral emotions. In R. J. Davidson, K. R. Scherer, and H. H. Goldsmith, editors, *Handbook of Affective Sciences*, pages 852—870. Oxford University Press, 2003.

[22] D. Harel, D. Kozen, and J. Tiuryn. *Dynamic Logic*. MIT Press, 2000.

[23] H. Hart. *Punishment and Responsibility: Essays in the Philosophy of Law.* Oxford University Press, Oxford, 1968.

[24] H. L. A. Hart and A. M. Honoré. *Causation in the law*. Clarendon, Oxford, 1959.

[25] A. Herzig and F. Schwarzentruber. Properties of logics of individual and group agency. *Advances in modal logic*, 7:133–149, 2008.

[26] A. Herzig and N. Troquard. Knowing how to play: uniform choices in logics of agency. In *Proceedings of AAMAS 2006*, pages 209–216. ACM, 2006.

[27] J. F. Horty. *Agency and Deontic Logic*. Oxford University Press, Oxford, 2001.

[28] J. F. Horty and N. Belnap. The deliberative STIT: A study of action, omission, and obligation. *Journal of Philosophical Logic*, 24(6):583–644, 1995.

[29] D. Kahneman and A. Tversky. The psychology of preferences. *Scientific American*, 246:160–173, 1982.

[30] B. Kooi and A. Tamminga. Moral conflicts between groups of agents. *Journal of Philosophical Logic*, 37(1):1–21, 2008.

[31] M. Lewis. Self-conscious emotions: Embarrassment, pride, shame, and guilt. In M. Lewis, J. Haviland-Jones, and L. F. Barrett, editors, *Handbook of emotions*, pages 742–756. Guilford Press, 2008.

[32] T. D. Lima, L. M. M. Royakkers, and F. Dignum. A logic for reasoning about responsibility. *Logic Journal of the IGPL*, 18(1):99–117, 2010.

[33] E. Lorini. Mutual Beliefs. In B. Kaldis, editor, *The Encyclopedia of Philosophy and the Social Sciences*. SAGE Publications, 2013.

[34] E. Lorini, D. Longin, B. Gaudou, and A. Herzig. The logic of acceptance: grounding institutions on agents' attitudes. *Journal of Logic and Computation*, 19(6):90140, 2009.

[35] E. Lorini and F. Schwarzentruber. A logic for reasoning about counterfactual emotions. *Artificial Intelligence*, 175(3-4):814–847, 2011.

[36] M. Miceli and C. Castelfranchi. How to silence one's conscience: Cognitive defenses against the feeling of guilt. *Journal for the Theory of Social Behaviour*, 28(3):287–318, 1998.

[37] A. Ortony, G. Clore, and A. Collins. *The cognitive structure of emotions*. Cambridge University Press, Cambridge, MA, 1988.

[38] P. Pettit. Responsibility incorporated. *Ethics*, January:171–201, 2007.

[39] B. Steunebrink. *The logical structure of emotions*. PhD thesis, Utrecht University, 2010.

[40] R. Tuomela. Group beliefs. *Synthese*, 91:28518, 1992.

[41] R. Tuomela. *The philosophy of social practices: A collective acceptance view*. Cambridge University Press, 2002.

[42] P. Turrini, J.-J. Meyer, and C. Castelfranchi. Coping with shame and sense of guilt: a dynamic logic account. *Journal of Autonomous Agents and Multi-Agent Systems*, 20(3):401–420, 2010.

[43] J. van Benthem. In praise of strategies. In J. van Eijck and R. Verbrugge, editors, *Discourses on social software*, Texts in Logic and Games. Amsterdam University Press, 2009.

[44] M. Y. Vardi. The taming of converse: Reasoning about two-way computations. In *Proceedings of the Conference on Logic of Programs*, volume 193 of *Lecture Notes in Computer Science*. Springer, 1985.

[45] H. Wansing. Tableaux for multi-agent deliberative-STIT logic. In G. Governatori, I. Hodkinson, and Y. Venema, editors, *Advances in Modal Logic, Volume 6*, pages 503–520. King's College Publications, 2006.

[46] G. Watson. Two faces of responsibility. *Philosophical Topics*, 24:227–248, 1996.

[47] B. Weiner. *Judgments of responsibility: A foundation for a theory of social conduct*. Guilford Press, 1995.