

# A Logic of Delegation

Timothy J. Norman<sup>a</sup>

<sup>a</sup>*Department of Computing Science,  
University of Aberdeen, Aberdeen AB24 3UE, Scotland UK*

Chris Reed<sup>b</sup>

<sup>b</sup>*School of Computing,  
University of Dundee, Dundee DD1 4HN, Scotland UK*

---

## Abstract

Delegation is a foundational concept for understanding and engineering systems that interact and execute tasks autonomously. By extending recent work on tensed action logic, it becomes possible to pin down a specific interpretation of responsibility with a well specified semantics and a convenient and intuitive logic for expression. Once descriptions of direct agent responsibility can be formed, there is a foundation upon which to characterise the dynamics of how responsibility can be acquired, transferred and discharged and, in particular, how delegation can be effected. The resulting logic, designed specifically to cater for responsibility and delegation, can then be employed to offer an axiological and semantic exploration of the related concepts of forbearance, imperatives and group communication.

*Key words:* Delegation, Groups, Imperatives, Responsibility, Agent Communication

---

## 1 Introduction

Delegation is a concept that pervades agent-based computing — tasks such as the purchase of goods within an electronic institution may be delegated to a software agent acting on behalf of a user [14,16], a goal may be delegated from one agent to another each representing different commercial organisations via,

---

*Email addresses:* [t.j.norman@abdn.ac.uk](mailto:t.j.norman@abdn.ac.uk) (Timothy J. Norman),  
[chris@computing.dundee.ac.uk](mailto:chris@computing.dundee.ac.uk) (Chris Reed).

for example, the Contract Net protocol [49], or an agent representing a manager within one organisation may, by virtue of their organisational position, delegate a task to an agent representing a subordinate [33]. Understanding the nature and complexities of delegation, therefore, has the potential to impact upon a wide audience within AI. In this paper we address the problem of understanding delegation directly by presenting a responsibility-based semantics of delegation (Section 2), that provides for an axiological account of how responsibility is transferred during delegation. With the basic mechanics of delegation in place, several tricky cases are then examined in detail, including:

- (1) the interaction between delegation and time (and specifically, how axioms of delegation might be extended in a tensed logic) (Section 3.1);
- (2) the relationship between responsibility and the concept of forbearance (and in particular, whether the Refref conjecture [2], by which activity is claimed to be equivalent to refraining from refraining, is a defensible axiom) (Section 3.2); and
- (3) the way in which group-addressed communication can effect delegative transfer of responsibility (and specifically, how distributive and collective responsibility [43] is composed from the responsibilities of the individuals in a group) (Section 3.3).

Finally, with a rich account of delegation in place, we explore how imperative communication can be used to execute delegation, how responsibility is acquired as a result, and how it can be discharged through appropriate action meeting the constraints of whole-hearted satisfaction [23] (Section 4).

## 2 Foundation

The first step in characterising delegation is to construct a model of agent responsibility, so that the former can be defined as a special case of the latter in which particular actions (often communicative actions) lead to a transfer of responsibility. Such a model needs to tie together agent intentions, actions, and states of the world. Elsewhere [42] we have argued that to handle this richness competently, it is necessary to adopt an approach that represents both states and events as first class objects. The argument there, and subsequent exploration of the system that results, leans heavily on foundational work carried out by Hamblin [23] in his investigation of imperatives. In this section we summarise a logic designed to capture the nature of the imperative based upon Hamblin’s Action-State Semantics. The logic captures, at both the semantic and syntactic levels, the important ontological distinction between ‘responsibility for the achievement of a state of affairs’ (captured by the modality *S*) and ‘responsibility for the execution of an action’ (captured by

the modality  $\top$ ). This is one of the key distinguishing features of the language, which we refer to as  $ST$ .

## 2.1 Syntax

In presenting the syntax of our language,  $ST$ , we begin by defining the set of well-formed formulae, then briefly discuss axioms and rules of inference of the modalities  $S$  and  $T$  and of a standard Peircian tense logic which is used to enable the expression of tensed responsibility formulae. This, along with the Hamblinian semantics of  $ST$  summarised in Section 2.2, lays the groundwork for the detailed analysis of the nature of delegation presented in section 3.

Basic atoms of the language  $ST$  include states of affairs, referred to using upper case Roman letters ( $A, B, C, \dots$ ), actions, which are referred to using lower case Greek letters ( $\alpha, \beta, \gamma, \dots$ ), and agents, for which we use  $x, y, \dots$ . We denote that a specific agent,  $x$ , executes action  $\alpha$  in the following manner:  $\alpha^x$ . Where the agent of an action is not specified, it is assumed that the action is carried out by some agent but it is not important which one.

The modalities  $S$  and  $T$  are relativised to specific agents and refer to state formulae and event formulae respectively. In this way,  $S_x A$  refers to agent  $x$  being responsible for the achievement of the state of affairs  $A$ , and  $T_x \alpha$  refers to  $x$  being responsible for the execution of action  $\alpha$ . Note that these modal statements do not specify any particular action for agent  $x$ . In satisfying  $T_x \alpha$ , for example, agent  $x$  may order some other agent,  $y$ , to carry out  $\alpha$ .

Any sentence in the language may be tensed through the use of the modalities  $G$  (always true in all futures) and  $H$  (always true in all pasts), and their respective duals,  $F$  (true at some point in a possible future) and  $P$  (true at some point in a possible past). Tensed sentences are  $S$ -formulae; to say that something will be true, or that some action has been done, etc. is a state of affairs. It is, however, entirely reasonable to permit tense operators to range over both states and events: states of affairs may have held in the past, and events may happen in the future, etc.

We may now define the well-formed formulae of our language  $ST$ . The basic atoms of our language are divided into two classes: (i) event formulae — those that consist entirely of propositional expressions of action (bound or unbound), and (ii) state formulae — all others. All such basic atoms are *wffs*. By conventional Propositional Logic ( $PL$ ), for any two *wffs*,  $\phi$  and  $\psi$ , that are event formulae,  $\phi \vee \psi$ ,  $\phi \wedge \psi$ ,  $\phi \rightarrow \psi$  and  $\neg\phi$  are also *wffs* that are event formulae. Similarly for any two *wffs* that are state formulae, any  $PL$  combination of them is also a *wff* that is a state formula. For the action modalities, any *wffs* that are event formulae can be used to form a further *wff*

with the  $\mathsf{T}$  modality:  $\mathsf{T}_x\alpha$ ,  $\mathsf{T}_x\alpha^x$ ,  $\mathsf{T}_x(\alpha^y \vee \beta)$ , etc., which are themselves state formulae. Any *wff* that is a state formula can be used to form a further *wff* with the  $\mathsf{S}$  modality  $\mathsf{S}_xA$ ,  $\mathsf{S}_x\mathsf{T}_y\alpha^z$ , etc. that are state formulae. Finally, for any  $\psi$  that is either an event formula or a state formula,  $\mathsf{G}\psi$ ,  $\mathsf{H}\psi$ ,  $\mathsf{F}\psi$  and  $\mathsf{P}\psi$  are also *wffs* that are state formulae.

The logic of modality  $\mathsf{S}_x$  is that of a regular modal system of type RT [10, p. 237]. This is the smallest system containing all axioms of Propositional Logic and closed under the rule of inference RE:

$$\text{RE} \quad \frac{A \leftrightarrow B}{\mathsf{S}_xA \leftrightarrow \mathsf{S}_xB}$$

with the additional axiom T, which is characteristic of logic of successful action

$$\text{T} \quad \mathsf{S}_xA \rightarrow A$$

and, of course, the distribution axiom K, which is minimally true of all modal logics

$$\text{K} \quad \mathsf{S}_x(A \rightarrow B) \rightarrow (\mathsf{S}_xA \rightarrow \mathsf{S}_xB)$$

Unlike other models of agentive action [3,11,26] however, we include neither the rule of necessitation ( $A/\mathsf{S}_xA$ ) nor that of anti-necessitation ( $\neg A/\neg\mathsf{S}_xA$ ). Consequently the logic of  $\mathsf{S}_x$  is non-normal. A key advantage of including neither of these axioms is that we may include the equivalence R without introducing inconsistency. Axiom R captures the intuition that if an agent is responsible for achieving  $A$  and responsible for achieving  $B$  then it is responsible for achieving the conjunction of these states, and that if an agent is responsible for achieving the conjunction of  $A$  and  $B$  then it is responsible for achieving each conjunct. In the logic of action specified by Jones and Sergot [26], for example, including axiom  $\mathsf{S}_x(A \wedge B) \rightarrow \mathsf{S}_xA \wedge \mathsf{S}_xB$  along with the rule of anti-necessitation will lead to a contradiction. Axiom R for modality  $\mathsf{S}_x$  is:

$$\text{R} \quad \mathsf{S}_x(A \wedge B) \leftrightarrow \mathsf{S}_xA \wedge \mathsf{S}_xB$$

Finally, we include axiom D in the logic of  $\mathsf{S}_x$ , which captures the intuition that an agent cannot be responsible for the achievement of contradictory states of affairs

$$\text{D} \quad \mathsf{S}_xA \rightarrow \neg\mathsf{S}_x\neg A$$

$$\begin{array}{lll}
\mathbf{G}\phi \rightarrow \psi \rightarrow (\mathbf{G}\phi \rightarrow \mathbf{G}\psi) & \mathbf{G}\phi \rightarrow \mathbf{F}\phi & \mathbf{G}\phi \rightarrow \mathbf{G}\mathbf{G}\phi \\
\mathbf{H}\phi \rightarrow \psi \rightarrow (\mathbf{H}\phi \rightarrow \mathbf{H}\psi) & \mathbf{H}\phi \rightarrow \mathbf{P}\phi & \mathbf{H}\phi \rightarrow \mathbf{H}\mathbf{H}\phi \\
\phi \rightarrow (\mathbf{H}\phi \rightarrow (\mathbf{G}\phi \rightarrow \mathbf{G}\mathbf{H}\phi)) & & 
\end{array}$$

Fig. 1. Axiom schemas of a Peircean tense logic.

The characterisation of modality  $\mathbf{T}_x$  is identical to that of  $\mathbf{S}_x$ ; both being regular modal systems of type RT. It should be noted, however, that these modalities operate over different worlds in their interpretation (see Section 2.2).

To enable us to explore the interpretation of responsibility over time, we require the use of a logic of time. For our purposes we adopt a simple Peircean tense logic. In this logic of time, there are two basic modalities:  $\mathbf{G}$  and  $\mathbf{H}$ . Their duals with respect to some sentence  $\phi$ , are defined as follows:  $\mathbf{F}\phi \stackrel{\text{def}}{=} \neg\mathbf{G}\neg\phi$  and  $\mathbf{P}\phi \stackrel{\text{def}}{=} \neg\mathbf{H}\neg\phi$ . The axioms for this tense logic are summarised in Fig. 1 (the standard Peircean model).

The well-formed formulae, rules of inference, definitions and axioms summarised in this section form the syntactic basis for our language ST. Before exploring the nature of delegation and the scope of an agent’s responsibility we outline the semantics of our language. Many readers may happily skip the following section, having an intuition of the meaning of the formulae of our language and wishing to ‘cut to the chase’. Those expecting a more thorough exploration of the semantics are referred to [42].

## 2.2 Semantics

The first step is to preserve the distinction between states and events simply by dividing them into separate sets of possible worlds. That is, some worlds contain state descriptions, and other worlds contain event descriptions. Then we define an accessibility relation that holds between these worlds. Rather than adopting a conventional binary relation, Hamblin’s semantics demands a ternary relation that links a world of state descriptions, with a world of event descriptions and another world of state descriptions.<sup>1</sup> In the following, we present the formal semantics for ST built upon a Hamblinian frame of reference,  $\mathcal{F}_{\mathcal{H}}$ .

**Definition 1** ( $\mathcal{F}_{\mathcal{H}}$ )  $\mathcal{F}_{\mathcal{H}} = \langle \mathcal{W}, \mathcal{R}_{\mathcal{H}} \rangle$  such that

(i)  $\mathcal{W}$  is a non-empty set that collects together our ‘state worlds’ and ‘event

<sup>1</sup> It could equally well be linking two event description worlds with one state description worlds; ultimately nothing of importance hangs upon this decision.

- worlds'; and
- (ii)  $\mathcal{R}_{\mathcal{H}}$  is a ternary relation, where  $\langle u, v, w \rangle \in \mathcal{R}_{\mathcal{H}}$  and  $u, v, w \in \mathcal{W}$ . Each  $\langle u, v, w \rangle \in \mathcal{R}_{\mathcal{H}}$  should be understood as 'state  $w$  is accessible from state  $u$  by way of  $v$ '.

Note that it is not necessary to enforce types upon the worlds explicitly in the semantics; this can be handled implicitly by  $\mathcal{R}_{\mathcal{H}}$ . There are two conditions, however, on  $\mathcal{R}_{\mathcal{H}}$  that are required to accurately capture the Hamblinian picture and the intuitions associated with it. The first is that each world in  $\mathcal{W}$  is either a world containing state descriptions (in this case it is a member of the set  $\mathcal{W}_{\text{states}}$ ) or a world containing event descriptions (in which it is a member of the set  $\mathcal{W}_{\text{events}}$  as defined above), but not both; i.e.  $\mathcal{W}_{\text{states}} \cap \mathcal{W}_{\text{events}} = \emptyset$ . The second constraint on our relation  $\mathcal{R}_{\mathcal{H}}$  is that the structure of interconnected 'state worlds' (via 'event worlds') is a directed acyclic graph. Beyond these two conditions, the relation  $\mathcal{R}_{\mathcal{H}}$  does not portray any of the more typical characteristics of many binary accessibility relations used by classical modal logics: it is neither symmetric nor reflexive, neither transitive nor Euclidean.

**Definition 2** ( $\mathcal{W}_{\text{states}}$ ) *Each element of  $\mathcal{R}_{\mathcal{H}}$  contains two state worlds (those appearing in the first and third places in the tuple).*

$$\mathcal{W}_{\text{states}} = \{w \mid \langle w, x, y \rangle \in \mathcal{R}_{\mathcal{H}}\} \cup \{w \mid \langle x, y, w \rangle \in \mathcal{R}_{\mathcal{H}}\}$$

**Definition 3** ( $\mathcal{W}_{\text{events}}$ ) *Each element of  $\mathcal{R}_{\mathcal{H}}$  contains a single event world (appearing in the second place in the tuple).*

$$\mathcal{W}_{\text{events}} = \{w \mid \langle x, w, y \rangle \in \mathcal{R}_{\mathcal{H}}\}$$

The semantic structure defined by  $\mathcal{R}_{\mathcal{H}}$  forms the lower layer in our model, and, on the basis of this layer, it is possible to define accessibility relations (and later, necessitation functions). First, for the temporal component we define an accessibility relation  $R_F$ , expressing the earlier-later relation.

**Definition 4** ( $R_F$ ) *An event world  $v$  occurs later than the state world  $u$  immediately preceding it (i), consecutive state worlds are related (ii), and, in (iii), transitivity is built into the relation.*

$$\langle u, v \rangle \in R_F \text{ iff } \exists w \in \mathcal{W} \text{ s.t. } \langle u, v, w \rangle \in \mathcal{R}_{\mathcal{H}} \text{ or} \quad (\text{i})$$

$$\exists w \in \mathcal{W} \text{ s.t. } \langle u, w, v \rangle \in \mathcal{R}_{\mathcal{H}} \text{ or} \quad (\text{ii})$$

$$\exists w, t \in \mathcal{W} \text{ s.t. } (\langle u, w, t \rangle \in \mathcal{R}_{\mathcal{H}} \text{ and } \langle t, v \rangle \in R_F) \quad (\text{iii})$$

To provide a foundation for temporal statements in our language we cannot simply express a later-earlier relation in terms on  $R_F$  (the earlier-later relation). This is due to our use of the ternary relation  $\mathcal{R}_{\mathcal{H}}$  and that all tensed sentences ( $H\psi$ ,  $F\psi$ , etc.) are state formulae. Our later-earlier relation,  $R_P$ , is, therefore, defined separately, and, together,  $R_F$  and  $R_P$  form a conventional temporal frame for a traditional, transitive tense logic [40].

**Definition 5** ( $R_P$ ) *The definition of  $R_P$  is a direct analog of  $R_F$ .*

$$\langle u, v \rangle \in R_P \text{ iff } \exists w \in \mathcal{W} \text{ s.t. } \langle w, v, u \rangle \in \mathcal{R}_{\mathcal{H}} \text{ or} \quad (\text{i})$$

$$\exists w \in \mathcal{W} \text{ s.t. } \langle v, w, u \rangle \in \mathcal{R}_{\mathcal{H}} \text{ or} \quad (\text{ii})$$

$$\exists w, t \in \mathcal{W} \text{ s.t. } (\langle t, w, u \rangle \in \mathcal{R}_{\mathcal{H}} \text{ and } \langle t, v \rangle \in R_P) \quad (\text{iii})$$

The action component is characterised in a slightly different way. Given that the logic of  $\mathbf{S}$  and  $\mathbf{T}$  is non-normal, it demands a minimal model, defined upon necessitation functions. Those necessitation functions must act upon a different substrate: for the  $\mathbf{S}$  modality, the substrate is state descriptions, for  $\mathbf{T}$ , event descriptions. The necessitation functions are relativised to individual agents in the usual way (that is, the way in which one agent's behaviour is described is independent of how other agents' behaviour is described). Thus  $\mathcal{S}^x$  is the necessitation function for the modality  $\mathbf{S}$ , relativised to some agent,  $x$ . The functions map from worlds to sets of worlds. So,  $\mathcal{S}^x : \mathcal{W}_{\text{states}} \rightarrow \wp(\wp(\mathcal{W}_{\text{states}}))$ , as usual (thereby picking out worlds by which necessity is defined). The  $\mathbf{T}$  modality is different in that  $\mathcal{T}^x : \mathcal{W}_{\text{states}} \rightarrow \wp(\wp(\mathcal{W}_{\text{events}}))$ ;  $\mathcal{T}^x$  is, therefore, picking out particular events that are, loosely, "actionable" by  $x$  from a state world  $w$ . Furthermore, the  $\mathcal{W}_{\text{events}}$  worlds are not simply propositional. To accurately model Hamblin's conception of "deed-agent assignments", these worlds are filled (exclusively) with statements of the form *agent  $x$  performs action  $\alpha$* , that we represent with the typographic shorthand  $\alpha^x$ , and *wffs* constructed from such statements using *PL*.

In this way, the model as a whole is defined as  $\langle \mathcal{W}, \mathcal{X}, \mathcal{I}, \mathcal{R}_{\mathcal{H}}, \mathcal{S}^x, \mathcal{T}^x \rangle$  for a set of possible worlds  $\mathcal{W}$ , a set of agents  $\mathcal{X}$ , an interpretation function  $\mathcal{I}$ , the ternary Hamblinian accessibility relation,  $\mathcal{R}_{\mathcal{H}}$  and the relativised necessitation functions for the modalities  $\mathbf{S}$  and  $\mathbf{T}$ ,  $\mathcal{S}^x$  and  $\mathcal{T}^x$  (for each  $x \in \mathcal{X}$ ), respectively.

The necessitation functions, in combination with the accessibility relations then offer a straightforward way of characterising the semantics of the logic as a whole:

$$\begin{aligned}
\models_{\omega}^{\mathcal{M}} A &\text{ iff } \mathcal{I}(A, \omega) = \top \\
\models_{\omega}^{\mathcal{M}} \alpha^x &\text{ iff } \mathcal{I}(\alpha^x, \omega) = \top \\
\models_{\omega}^{\mathcal{M}} \alpha &\text{ iff } \exists x \in \mathcal{X} \text{ s.t. } \mathcal{I}(\alpha^x, \omega) = \top \\
\models_{\omega}^{\mathcal{M}} \mathsf{S}_x A &\text{ iff } \|A\|^{\mathcal{M}} \in \mathcal{S}^x(\omega) \\
\models_{\omega}^{\mathcal{M}} \mathsf{T}_x \alpha &\text{ iff } \|\alpha\|^{\mathcal{M}} \in \mathcal{T}^x(\omega)
\end{aligned}$$

The truth set is constructed normally:

$$\|\varphi\| = \{\omega \mid \models_{\omega}^{\mathcal{M}} \varphi\}$$

The truth set is constructed in the same manner for both states and events; this symmetry is a result of the typing of possible worlds, so that increased complexity in the model structure yields increased simplicity in the connection between that structure and the syntactic surface.

Now we are in a position to characterise the tense component of our language. Here, we are simply defining sentences with our tense modality with accessible worlds in which the object of that modality are true. For example, we can say in some world,  $\omega$ , given some model,  $\mathcal{M}$ , that  $\phi$  holds in some point in a possible future,  $\mathsf{F}\phi$ , if and only if there is some other world,  $\omega'$  that is accessible from  $\omega$  via our future-directed relation,  $R_F$ , in which  $\phi$  holds.

$$\begin{aligned}
\models_{\omega}^{\mathcal{M}} \mathsf{F}\phi &\text{ iff } \omega \in \mathcal{W}_{\text{states}} \wedge \exists \omega' \in \mathcal{W} \text{ s.t. } R_F(\omega, \omega') \text{ and } \models_{\omega'}^{\mathcal{M}} \phi \\
\models_{\omega}^{\mathcal{M}} \mathsf{G}\phi &\text{ iff } \omega \in \mathcal{W}_{\text{states}} \wedge \forall \omega' \in \mathcal{W} \text{ s.t. } R_F(\omega, \omega') \text{ and } \models_{\omega'}^{\mathcal{M}} \phi \\
\models_{\omega}^{\mathcal{M}} \mathsf{P}\phi &\text{ iff } \omega \in \mathcal{W}_{\text{states}} \wedge \exists \omega' \in \mathcal{W} \text{ s.t. } R_P(\omega, \omega') \text{ and } \models_{\omega'}^{\mathcal{M}} \phi \\
\models_{\omega}^{\mathcal{M}} \mathsf{H}\phi &\text{ iff } \omega \in \mathcal{W}_{\text{states}} \wedge \forall \omega' \in \mathcal{W} \text{ s.t. } R_P(\omega, \omega') \text{ and } \models_{\omega'}^{\mathcal{M}} \phi
\end{aligned}$$

### 3 Delegation and Responsibility

With this formal characterisation of the modalities  $\mathsf{S}$  and  $\mathsf{T}$  in place, we now explore the nature of delegation and, in doing so, develop a detailed model of the scope of an agent's responsibility. Delimiting the scope of agentive responsibility in this way provides a foundation for some of the key questions in the design, development and operation of multi-agent systems. It forms an essential component for performance monitoring, checking compliance with protocols, and identifying the contributors to a failure which may impact upon the reputations of agents within a society.

Fundamental to our notion of responsibility, of course, is the semantics of



the modalities S and T, which captures our conception that an agent be held responsible for its own volition. In broader terms, however, we must determine to what extent an agent can be held responsible with respect to delegated activities. We start this discussion with a summary of the basic axioms of delegation for our logic containing the action modalities S and T and tense modalities F and P (and hence their duals G and H respectively), which were first outlined in [42]. These lay the foundation for an extended analysis in which we explore the notion of refraining and group-directed imperatives. In the subsequent three sections, therefore, we focus our attention on the following questions:

- (1) If an agent acts through another (i.e. if an agent delegates a task to another), should it be deemed to have acted in person? (Section 3.1.)
- (2) Is *doing* the same as *refraining from refraining*? (Section 3.2.)
- (3) What does it mean for an imperative to be issued to a group of agents? (Section 3.3.)

### 3.1 Responsibility for Delegated Tasks

In common with the model of agentive action proposed by Chellas [11], but contrary to Von Wright’s [52] characterisation and that of Belnap *et al.* [4], the theory presented here offers scope for nesting the two modalities in building a rich notion of responsibility. Thus, in this section we discuss the theorems and axioms of delegation summarised in Fig. 2.

An agent may use many means to delegate tasks; e.g. the issuing of a command within the context of a military organisation, or asking a colleague to cover for a lecture. As a result, the agent having delegated a task will have some responsibility for the task delegated, but what is the nature of this responsibility and how can it be expressed? To answer this question, we must explore the meaning of formulae in which the S and T modalities are nested. For example, if the activity concerned is the performance of action  $\alpha$ , the delegator is agent  $x$  and the agent to whom responsibility for the act is delegated is  $y$ , this can be captured by the sentence  $S_x T_y \alpha$ . Similarly, if the activity concerned is the achievement of some state of affairs,  $A$ , this can be captured by the sentence  $S_x S_y A$ . But, what do these sentences mean and where lies the responsibility for, respectively,  $\alpha$  and  $A$ ?

The first thing to notice is that specialisations of axiom T give us part of the answer. In Fig. 2, theorems TSS and TST are presented. Taking TST for instance, if it can be said that agent  $x$  is responsible for agent  $y$ ’s responsibility for the performance of action  $\alpha$ , then we can conclude that  $y$  is indeed responsible for the performance of  $\alpha$ . But what of  $x$ ’s responsibilities?

TSS	$S_x S_y A \rightarrow S_y A$	QS	$S_x S_y A \rightarrow S_x A$
TST	$S_x T_y \alpha \rightarrow T_y \alpha$	QT	$S_x T_y \alpha \rightarrow T_x \alpha$
TFS	$FS_x A \rightarrow FA$	QFS	$S_x FS_y A \rightarrow S_x FA$
TPS	$PS_x A \rightarrow PA$	QPS	$S_x PS_y A \rightarrow S_x PA$
TFT	$FT_x \alpha \rightarrow F\alpha$	QFT	$S_x FT_y \alpha \rightarrow S_x F\alpha$
TPT	$PT_x \alpha \rightarrow P\alpha$	QPT	$S_x PT_y \alpha \rightarrow S_x P\alpha$

Fig. 2. Theorems and Axioms of Delegation and Responsibility.

Chellas [11] provides us with an answer in the form of the legal principle *Qui facit per alium facit per se* (roughly, he who acts through another has acted himself). This principle led to Chellas's introduction of an axiom Q; equivalents of this concept are presented in Fig. 2 as QS and QT. Axiom QS expresses the idea that if it is the case that agent  $x$  is responsible for agent  $y$ 's responsibility for the achievement of  $A$ , then it can also be said that  $x$  is responsible for the achievement of  $A$ . It is interesting to note that these axioms cannot be accepted by Belnap and Perloff [3]: Chellas [11, p. 506] shows that the *something happens* condition (for all possible combinations of choices of all agents there is at least one history) leads to (using our notation)  $S_x S_y A$  being false whenever  $x \neq y$ .

Axioms QS and QT capture the essence of our answer to the first question posed in the introduction to this section, but this requires some refinement when the issue of tense is considered. In order to fully explore other possible theorems and axioms, we must consider all 2- and 3-modality well-formed formulae with alternative tense and action modalities. Specifically, we must consider the meaning of all formulae that can be composed from the following schemes (a), (b), (c) and (d):

$$(a) S_x \frac{F A}{P \alpha} \quad (b) \frac{F S_x A}{P T_x \alpha} \quad (c) S_x \frac{F S_y A}{P T_y \alpha} \quad (d) \frac{F}{P} S_x \frac{F A}{P \alpha}$$

These summarise a total of 20 possible combinations of 2- and 3-modality *wffs* where tense and action modalities alternate. In these schemes the horizontal lines indicate that one of the two modalities above or below the line may be included to form a *wff*. Using scheme (c), for example, we can construct formulae commencing with the modality  $S_x$ , followed by either F (true at some point in a possible future) or P (true at some point in a possible past), and then followed by either  $S_y A$  or  $T_y \alpha$ . The formula  $S_x FS_y A$  can, in this way, be constructed from this scheme, which expresses the notion that agent  $x$  is responsible for it being the case at some point in a possible future that  $y$  is responsible for the achievement of  $A$ . Similarly,  $S_x FT_y \alpha$  and  $S_x PS_y A$  can be constructed from this scheme.

The entailments derived from formulae that can be created from scheme (a) are

relatively straightforward. The T axiom will, in each case, allow the derivation of a tensed formula; e.g. from  $S_xFA$  we may derive  $FA$  and from  $S_xP\alpha$  we may derive  $P\alpha$ . There are no further axioms required; we would certainly not wish to derive  $S_xA$  or  $FS_xA$  from  $S_xFA$  for example.

Next, let us consider those *wffs* that can be created from scheme (b). In order for us to derive appropriate conclusions from these *wffs* we do require further axioms. In fact, we need analogues of the T axiom to retain the notion of a model of successful action. We introduce the four further axioms, TFS, TPS, TFT and TPT (Fig. 2), for this purpose. These analogues of the T axiom ensure that the tense (loosely) associated with an action is carried over to the successful completion of that action. Thus from  $FS_xA$  (it is true at some point in a possible future that agent  $x$  is responsible for the achievement of  $A$ ), we want to conclude  $FA$  ( $A$  is true at some point in a possible future), from  $PT_x\alpha$  we want to conclude  $P\alpha$ , and so on.

Schemes (c) and (d) provide *wffs* that include three alternating tense and action modalities. In the case of the formulae that may be created from scheme (d), the existing axioms are sufficient (including the tensed axioms of successful action introduced when discussing scheme (b)). For example, from  $FS_xFA$  we can, through TFS, derive  $FFA$  and hence  $FA$ . Therefore, all the deductions one would wish to be able to draw are already catered for.

Formulae that may be created from scheme (c), however, require further axioms to fully characterise delegation of responsibility over time. Thus  $S_xFS_yA$  yields not only  $FS_yA$  via axiom T, and thence  $FA$  via axiom TFS, but in addition, we want to capture the fact that  $S_xFS_yA$  also has a more intimate connection (i.e.  $x$ 's responsibility for) the future occurrence of  $A$ . Specifically, by analogy to the atemporal Q axioms, we would want to be able to derive  $S_xFA$  (axiom QFS in Fig. 2).

In this way, we may construct four new analogs of the Q axiom (QFS, QPS, QFT and QPT, Fig. 2) that carry over tense modalities, in just the same way as we have done for analogs of the T axiom above. To illustrate the role of these axioms, let us consider just one of them and summarise its meaning. Axiom QFT captures the idea that *if* agent  $x$  is responsible for, at some point in a possible future, agent  $y$  ensuring that  $\alpha$  is done, *then*  $x$  is responsible for, at some point in a possible future,  $\alpha$  being done.

The meanings of axioms QPS and QPT are, possibly, less intuitively clear. It may be a little difficult to accept QPS in terms of our intuitions about time (and causality), but it should be noted that what we are capturing is the notion of agentive responsibility in a temporal context. Thus, QPT may be better understood using the following characterisation: if agent  $x$  is responsible for a state of affairs in which, at some point in a possible past, agent  $y$  ensures

1.  $S_{\text{consultant}}F_{\text{patient\_at\_home}}$  (patient released)
2.  $F_{\text{patient\_at\_home}} \leftrightarrow FS_{\text{patient}}\text{family\_murdered}$
3.  $S_{\text{consultant}}FS_{\text{patient}}\text{family\_murdered}$
4.  $S_{\text{consultant}}F_{\text{family\_murdered}}$  (by QFS)
5.  $FS_{\text{consultant}}\text{family\_murdered}$  (by QFS<sup>†</sup>)

Fig. 3. Consultant responsibility example.

that  $\alpha$  is done, then  $x$  is responsible for, at some point in a possible past, the doing of  $\alpha$ .

An alternative to axiom QFS is defensible: we might want to allow  $FS_xA$  to be derived from  $S_xFS_yA$ ; i.e. that by  $x$  being responsible for it being the case that it is possible in the future for agent  $y$  to be responsible for the achievement of  $A$ , it is possible in the future for  $x$  to be responsible for  $A$ . We label this plausible, but stronger, alternative QFS<sup>†</sup>:

$$\text{QFS}^\dagger \quad S_xFS_yA \rightarrow FS_xA$$

Similarly, axioms QPS<sup>†</sup> ( $S_xPS_yA \rightarrow PS_xA$ ), QFT<sup>†</sup> ( $S_xFT_y\alpha \rightarrow FT_x\alpha$ ), and QPT<sup>†</sup> ( $S_xPT_y\alpha \rightarrow PT_x\alpha$ ) are candidates for capturing a notion of responsibility in defining a social system. The question is why are these stronger delegation axioms, and what are the consequences of adopting these alternatives?

To illustrate why axiom QFS<sup>†</sup> introduces a stronger notion of responsibility (and, by analogy, the other ‘<sup>†</sup>’ axioms given above), consider the example of a mental health consultant making the decision to release a patient from a confined psychiatric ward. Now, suppose that the patient returns home and murders his family. Where lies the responsibility for this act? Formalising this scenario, we can say that the consultant is responsible for it being the case that it is possible in the future that the patient is at home (due to the consultant having released the patient); line 1 in Fig. 3. Let us suppose that the patient murdering his family is (future) possible if and only if it is (future) possible that the patient is at home; line 2 in Fig. 3. By rule RE, we can conclude line 3; i.e. that the consultant is responsible for it being a future possibility that that patient is responsible for killing his family. Now, lines 4 and 5 offer two possible conclusions that may be drawn from line 3; leading to two possible axiomatic formulations: one with the conclusion given on line 4 of Fig. 3 (through QFS in Fig. 2), and one with the conclusion given on line 5 of Fig. 3 (through the alternative axiom QFS<sup>†</sup>).

This emotive example generates strong intuitions. We may take the view that the mental patient is only partly responsible for his actions, and that the consultant must shoulder some of the responsibility. One might expect a charge of professional negligence to be levelled. But, intuitively, it is unlikely that the consultant would be seen to be guilty of (conspiracy to) murder or manslaughter. This is the choice offered by QFS and QFS<sup>†</sup>. Line 4 of Fig. 3 describes

responsibility of a possibility, in which culpability is only established through, for example, some assessment of fore-knowledge of the biconditional in line 2. Responsibility for possible disaster is a common feature in assessing negligence. Line 5, in contrast, is quite different, as it describes the possibility of direct responsibility; this being, in example, sufficient for graver charges being laid. Cases such as this seem to suggest strongly that it may not be reasonable to include  $\text{QFS}^\dagger$  as an axiom of delegation in many social systems, and we would argue that it is not generally considered valid in legal systems.

### 3.2 Forbearance

Belnap *et al.* [4], spend some time discussing and laying out the formal properties of what Belnap [2] describes as the Refref conjecture. Belnap describes Refref in such a clear and engaging way that we include here an extended quote:

The background idea is that one good way of approaching agency is via the modal locution  $[\alpha \text{ stit} : Q]$ , to be read as “ $\alpha$  sees to it that  $Q$ ”, where  $\alpha$  is an agent and where according to the thesis that the complement of stit should be unrestricted,  $Q$  may take the place of any sentence whatsoever.

Given the locution  $[\alpha \text{ stit} : Q]$ , it is easy to see that refraining when complemented by a *non*-agentive has just the form  $[\alpha \text{ stit} : \neg Q]$ , for example *Autumn Jane refrains from becoming muddy* comes to *Autumn Jane sees to it that she does not become muddy*. Accordingly, *refraining from an action* has the form  $[\alpha \text{ stit} : \neg[\alpha \text{ stit} : Q]]$ . For example, *Autumn Jane refrains from seeing to it that she becomes muddy* comes to *Autumn Jane sees to it that it is false that she sees to it she becomes muddy*. The two forms are easy for the ear to confuse, but the reflective eye can see that the advice to refrain from seeing to it that one becomes muddy is much easier to follow than advice to refrain from becoming muddy. Parents and children alike doubtless rely on the ear’s confusion when they hash out the matter with each other after the dress is splattered by a passing truck.

Give a modal logician a little nesting and more is wanted. The form  $[\alpha \text{ stit} : \neg[\alpha \text{ stit} : \neg[\alpha \text{ stit} : Q]]]$ , which may be read as  $\alpha$  refrains from refraining from seeing to it that  $Q$ , illustrates the nesting of refraining within itself. In this language, the sample question noted above can be expressed as the *Refref conjecture*.  $[\alpha \text{ stit} : \neg[\alpha \text{ stit} : \neg[\alpha \text{ stit} : Q]]]$  is equivalent to  $[\alpha \text{ stit} : Q]$ . If the Refref conjecture is true, then the only way that Autumn Jane can refrain from refraining from seeing to it that she becomes muddy is to see to it that she becomes muddy.

The conjecture is perhaps not so exciting in itself, though I confess to a certain mud-pie fondness for it.

In [4] the Refref conjecture is explored and shown to hold for both a-stit and d-stit (though the complexity of demonstrating the former far exceeds doing so for the latter). It might be expected from a purely formal point of view that there would be an analog in the logic of S thus:

$$S_x \neg S_x \neg S_x A \equiv S_x A$$

The logic of T might be expected to be a little different (nested T statements are, after all, not even *wffs*), but brief inspection yields the following candidate:

$$S_x \neg S_x \neg T_x \alpha \equiv T_x \alpha$$

Despite initial intuitions to the contrary, neither of these capture Refref. The problem arises from a superficial reading of Belnap [2] (and Pörn) [39], and in particular, from mis-associating the rightmost *stit* in Belnap’s formulation with an  $S_x$  or  $T_x$  expression. One indicator that something is awry is in Belnap’s description of “becoming muddy” as non-agentive: though it may be non-agentive, it is certainly action-oriented, even though the agent of the action is un-specified. (One cannot have becoming muddy as anything other than an action: even from a purely grammatical point of view, “becoming” cannot be acting as a gerund after “refrains”, so it must be the participle, i.e., a statement referring to activity.) The contrast he draws between that and the second formula is thus overstating the case: both  $[\alpha stit : \neg Q]$  and  $[\alpha stit : \neg[\alpha stit : Q]]$  are concerned with refraining from action, it is just that the first concerns an action in which  $\alpha$  is a hapless victim (such as the passing truck that splatters the dress), whereas the second concerns an action specifically of  $\alpha$ . Translation of Refref into the logics of S and T is thus not simply a matter of translating  $\alpha stit$  into  $S_\alpha$ . The closest we might come to such translation is to compare  $[\alpha stit : Q]$  with an action description such as  $Q^\alpha$ : the action of  $\alpha$  seeing to it that  $Q$  (in Belnap’s terms) is analogous to the deed of  $Q$  being performed by  $\alpha$  (in Hamblin’s terms). (We are not arguing for a translation that maps perfectly either syntactically or semantically. The aim is only to use the comparison as a stepping stone to expressing Refref). Unfortunately, though, for a simplistic attempt at persisting in this vein, an action description such as  $Q^\alpha$  cannot be further nested. The problem we are facing is that in Belnap’s account, responsibility for an action is exactly co-extensive with an agent’s performance of that action. In the logic of S and T, performance of an action is quite distinct from the responsibility for the execution of that action. Such a distinction is, of course, useful in capturing subtle issues in the law (such as the concept of “diminished responsibility”) and is invaluable in accounting for delegation. But it makes understanding

	Positive responsibility	Negative responsibility
Personal action $\alpha^x$	Responsibility for $x$ doing $\alpha$ <i>Performance</i> of $\alpha$ $T_x\alpha^x$	Responsibility for $x$ not doing $\alpha$ <i>Forbearance</i> from $\alpha$ $T_x\neg\alpha^x$
Action $\alpha$	Responsibility for $\alpha$ <i>Assurance</i> of $\alpha$ $T_x\alpha$	Responsibility for not $\alpha$ <i>Interdiction</i> of $\alpha$ $T_x\neg\alpha$
State $A$	Responsibility for $A$ <i>Establishment</i> of $A$ $S_xA$	Responsibility for not $A$ <i>Prohibition</i> of $A$ $S_x\neg A$

Table 1  
Types of expressions of responsibility

forbearance difficult: are we talking about forbearing from carrying out an action, or forbearing from taking on the responsibility?

There are strong linguistic clues as to what is going on. One talks of “forbearing to do something” or “refraining from some activity”. That is, both forbearing (in this sense) and refraining predicate action. The logic of S and T, of course, is equipped to express and distinguish action and responsibility for action. One agent,  $x$ , simply happening to not perform some action  $\alpha$  themselves is expressible as  $\neg\alpha^x$ . But as Belnap and Pörn have pointed out, this is not enough to capture the responsibility involved in forbearing. For that, we need  $T_x\neg\alpha^x$ :  $x$ ’s (T-) responsibility for  $x$  not doing  $\alpha$ . The non-agentive (or, more precisely, the implicitly universally quantified agentive) responsibility in  $T_x\neg\alpha$  is much stronger, capturing  $x$ ’s responsibility for the non-execution of  $\alpha$  by any agent. This is capturing the notion of *interdicting*. Responsibility for a given state of affairs not pertaining — or what we might loosely call *prohibiting* — can be straightforwardly captured by a third case:  $S_x\neg A$ . States of affairs are intrinsically non-agentive, and so the catalogue ends at this point, except to note that for each of these expressions of "negative" responsibility, there is a dual that omits the negation (each of which has been encountered in Section 2). The six types are laid out in Table 1. (In describing negative responsibility, the relationship between the first and third cases could be emphasised terminologically by referring to  $S_x\neg A$  as *forfending*; similarly, that between the second and third cases could be emphasised by having  $T_x\neg\alpha$  as *proscribing*; the terms in the table avoid undue emphasis on either pairing).

In this way, we can usefully break down the concept of responsibility into six distinct types: performance, assurance and establishment, and their duals, forbearance, interdiction and prohibition. These terms are not ideal. Quite apart from the slight mismatches between their everyday meanings and the needs of the taxonomy here, there is a further terminological problem in that

the six all have their origins in verbs. Forbearing, interdicting and so on are all events, whereas these descriptions of responsibility are all states; for Belnap, forbearing is, of necessity, an action description, since his *stit* operator ranges only over action descriptions. Here, forbearance and its colleagues are treated strictly as species of responsibility, i.e. as state descriptions: the Table’s more long-winded expressions of “Responsibility for ...” are more accurate if less concise.

Where then does that leave the Refref conjecture in the logic of S and T? There are now three possible interpretations, depending upon how exactly one interprets Belnap’s concept of “refraining”. Given that both S and T expressions are themselves state expressions, the first (i.e. outer) Ref of Refref must be negative responsibility with respect to a state, or what we have called prohibition in the table above. For the second (i.e. inner) Ref, each of the three types of negative responsibility are defensible candidates, giving rise to the three possible interpretations. The first is that Refref is prohibition of forbearance, i.e.  $S_x \neg T_x \neg \alpha^x$ , in which case our candidate for the conjecture is

$$S_x \neg T_x \neg \alpha^x \equiv T_x \alpha^x$$

If we take *Autumn Jane jumps in the muddy puddle* and *Autumn Jane avoids the muddy puddle* as our  $\alpha^x$  and  $\neg \alpha^x$ , respectively, we can form a cumbersome accurate gloss thus: Autumn Jane is responsible for the state in which she is not responsible for herself avoiding the muddy puddle. More loosely, and more informatively, we have, Autumn Jane prohibits herself from forbearing to jump in the puddle. Culpability in this case strikes the ear quite clearly.

The second interpretation is that Refref is prohibition of interdiction. This offers

$$S_x \neg T_x \neg \alpha \equiv T_x \alpha$$

If this time we take the action  $\alpha$  to be the throwing of a mud pie at Autumn Jane, we have an accurate gloss thus: Autumn Jane is responsible for the state in which she is not responsible for the action of someone throwing a mud pie at her. More loosely, a gloss reads, Autumn Jane prohibits herself interdicting mud pie throwing. Her culpability seems here much harder to demonstrate. Can she be said to be responsible for the action of the mud pie throwing merely through her deciding not to interdict it? Intuition says not (assuming no other encouragement from Autumn Jane).

The third interpretation is that Refref is prohibition of prohibition. This yields the equivalence

$$S_x \neg S_x \neg A \equiv S_x A$$

With the state  $A$  as Autumn Jane being muddy, the gloss is that Autumn Jane is responsible for the state in which she is not responsible for being mud-free. More perspicuously, Autumn Jane is responsible for not prohibiting her



muddiness. This case is closely analogous to the previous. In both cases, Autumn Jane is working to avoid responsibility for staying clean (i.e. prohibiting her responsibility for cleanliness). In so doing though, she cannot be said to have taken responsibility for any subsequent muddiness. A further source of evidence that indicates that our intuitions hold up in the axiomatisation arises from  $\mathsf{T}$ , by which  $\mathsf{S}_x A$  yields  $A$  (i.e.  $\mathsf{S}$  and  $\mathsf{T}$  comprise a logic of successful action). Our intuitions suggest that neither  $\mathsf{S}_x \neg \mathsf{T}_x \neg \alpha$  nor  $\mathsf{S}_x \neg \mathsf{S}_x A \equiv \mathsf{S}_x A$  require  $\alpha$  or  $A$  to hold — merely that they might, and if they do it has nothing to do with Autumn Jane. In contrast, if she prohibits herself from forbearing to leap in the muddle, as in the first case, then we certainly expect her to do it — and to suffer whatever sanctions may then loom.

The terms, “prohibition” and “interdiction”, in particular, have strong normative and legalistic connotations, so it is important to emphasise that the intended reading — and the logical forms — are entirely based on the actuality of the Hamblinian semantic frame, rather than upon a deontic ideality built upon it. That is not to say that there are not strong resonances: Lindahl’s [30] one-agent types and Hart’s [24] protective perimeter of rights provide exactly the right level of detail for expressing whether or not an agent is permitted to forbear, and how other agents might impinge upon an agent that is prohibiting interdiction. The logic of  $\mathsf{S}$  and  $\mathsf{T}$  is designed with such normative accounts clearly in mind, but an exploration of their interrelations is beyond the scope of this paper.

The Refref conjecture is somewhat reminiscent of the equivalence  $\neg\neg P \equiv P$  in propositional logic; perhaps a critique might be expected to have an analogue in intuitionistic logic in which the equivalence is rejected. The reason for rejecting  $\neg\neg P \equiv P$  in intuitionistic logic is a consequence of the constructivist view of the Law of Excluded Middle: “the intuitionist reject[s] the platonistic notion of mathematical truth as obtaining independently of our capacity to give a proof.” [18, p. 18] The platonistic inevitability of  $A \vee \neg A$  is far less compelling when the disjuncts are ‘ $x$  does  $\alpha$ ’ and ‘ $x$  refrains from doing  $\alpha$ ’, so perhaps it is not surprising that a more fine-grained analysis is required for a logic that encompasses action and responsibility.

In conclusion, Belnap sees the Refref conjecture as a question: is it the case that the only way to refrain from refraining from something is to do that something? If Refref holds, Belnap claims, the answer is yes. But on the current account, there is more to be said: Refref is about getting one’s hands dirty. Mere responsibility, as captured by  $\mathsf{S}$  and  $\mathsf{T}$  is not enough for Refref to go through. The only way to prohibit one’s own forbearance is to execute the action oneself. But in the more general case, it is possible to avoid responsibility of responsibility avoidance without having to commit the action oneself — it may, instead, come about through delegation, knowledge of the world, or mere chance. One may wish to go on to build an account of, for example,

legal culpability based upon causal knowledge, intent, malice aforethought, or other arbitrarily complex and nested representations of agents and their interactions with their environment. But equipped with nothing more than a logic of action, it is possible to distinguish between the intuitively compelling direct involvement demanded by Refref for personal action, which forms a part of the logic of S and T, and is the motivation for accepting Refref in [4], and the much more suspect Refref for interdiction and prohibition which is rejected here, but cannot be disentangled and excluded in the account of [4].

### 3.3 *Group Delegation and Joint Responsibility*

Here, we extend our theoretical model of responsibility to groups rather than individuals, and specifically, to the case of an imperative being issued to a group. In other words, what does it mean for an agent to delegate (issue an imperative) to a group of agents, and what are the consequent responsibilities of each of the parties involved?

In analysing group-directed imperatives, and hence approaching answers to these questions, let us start with a few simple examples:

- (i) “All of you stand up!”
- (ii) “Someone shut the door!”
- (iii) “Form a circle!”
- (iv) “Flip the switch!”

The first two of these examples illustrate a key distinction between the group being addressed *distributively* and as a *collective* (the first reference to this distinction being, to our knowledge, in the insightful monologue by Rescher [43]). Rescher uses the terms “distributive” and “collective” groups; terms that are similarly adopted in this research [36]. In example (i), the group is being addressed distributively — each student should stand up — and in example (ii) the group is being addressed as a collective — at least one student (but possibly all) should close the door. The third and fourth examples are, at first glance, more troublesome. Intuitively, the request to form a circle may indicate all those addressed, but not necessarily; the imperative may very well be satisfied by five or six members of the group addressed forming a circle — the wording of the example is ambiguous — but it would be difficult to argue that a single agent would be sufficient.<sup>2</sup> The request to flip the switch is similarly challenging; given the assumption that flipping a switch changes the state of an environment variable from one to another, the number of times that the switch is flipped is important in the satisfaction of the imperative.

---

<sup>2</sup> A single wagon riding in a defensive circle is used as a gag in the 1974 Mel Brooks film *Blazing Saddles*, and is perhaps the exception that proves the rule here!

This particular example may be criticised on the basis of formulation — why not request that the light (or whatever environmental variable is being referred to) be on/off — but it may very well be the case that an imperative is issued to the effect that one and only one member of a group act in such and such a way. Or, for that matter, exactly two, or exactly three members of the group to whom the imperative is issued.

It is important to note here that we make no assumptions within the model of any social or organisational context or structure within the group to which an imperative is issued. In our view, the consideration of such structure (e.g. hierarchies/teams that may exist within the group addressed) is outside a theory of delegation and responsibility, simply because an imperative, issued to a (collective or distributive) group of agents, is issued to every member of that group regardless of any additional organisational structure. This does not mean that organisational roles and relationships do not give weight to the delegation of an activity; this is an essential part of the context in which activities are delegated (see Section 4). Consider, for example, the CEO of a company issuing a directive regarding the company policy on (self-)certification of illness. This is directed to all employees of the company regardless of their position. It may be that this imperative is issued through the distribution of a memo — the mode of delivery is not important — but it applies to all those to whom it is directed. Suppose, in contrast, that the CEO issues the imperative to her heads of department that they each reduce the costs of their department by 10% over the following financial year. This is directed only to the heads of department. The instruction may provide weight to (give some justification for) any subsequent imperative issued by a head of department regarding the reduction of costs within their department. This subsequent imperative is, however motivated by the first, not the simple transmission of the CEO’s imperative.

In this analysis, let us start with the definition of  $S_{\hat{X}}A$ , thus:

**Definition 6** ( $S_{\hat{X}}A$ ) *All agents (in the set  $X \subseteq \mathcal{X}$ ) are responsible for the achievement of the state of affairs  $A$ .*

$$S_{\hat{X}}A \stackrel{\text{def}}{=} \bigwedge_{x \in X} S_x A$$

The question that we will consider here is whether this is sufficient to capture our intuitions about imperatives with respect to the achievement of a state of affairs directed to a distributive group.

Suppose that  $X$  is {Alice, John}; by definition for that group to be responsible for the achievement of  $A$ , both Alice and John must see to it that  $A$ . Suppose Alice and John are requested to make sure that their timesheets for the

month are completed for tomorrow, the state  $R$ . In accordance with the above definition, this is equivalent to  $S_{\text{Alice}}R \wedge S_{\text{John}}R$ , and, if uttered in an appropriate context, Alice and John would be suitably motivated to whole-heartedly satisfy this imperative, and take necessary action.

Suppose that John fails to submit his timesheet. If Alice does submit her timesheet on schedule, can it be said that the original group-directed imperative has been whole-heartedly satisfied? Clearly not:  $S_{\text{Alice}}R$  is satisfied,  $S_{\text{John}}R$  is not. Thus, this definition does account for the idea that for the group-directed imperative to be satisfied, all those addressed must whole-heartedly satisfy the individual imperatives that flow from it. This is, however, a rather weak notion of group responsibility; Alice is not responsible for John doing his part (and hence cannot be considered, in part, culpable if he does not), and, similarly, John has no directive to ensure that Alice contributes. Although it captures some aspects of what it means for agents to satisfy a group activity, it does not require that they *act as a team*.

Consider an alternative, stronger, interpretation of the issuing of an imperative to a distributive group that emphasises group responsibility.

**Definition 7** ( $S_{[X]}A$ ) *Each member of the group  $X$  is responsible for each member being responsible for achieving the state of affairs  $A$ . (An alternative formulation of this definition in terms of  $S_{\hat{X}}$  is given in square brackets.)*

$$S_{[X]}A \stackrel{\text{def}}{=} \bigwedge_{x \in X} S_x \left( \bigwedge_{y \in X} S_y A \right) \quad \left[ \text{or } S_{[X]}A \stackrel{\text{def}}{=} S_{\hat{X}} S_{\hat{X}} A \right]$$

Now, using the same example as above, the imperative  $S_{\{\{\text{Alice}, \text{John}\}\}}R$  expands (reducing  $S_x S_x A$  to  $S_x A$  by T throughout) to:  $S_{\text{Alice}}R \wedge S_{\text{Alice}} S_{\text{John}}R \wedge S_{\text{John}}R \wedge S_{\text{John}} S_{\text{Alice}}R$ . This introduces a flavour of group responsibility for the satisfaction of the state of affairs concerned — as well as Alice and John being individually responsible for the submission of their own timesheets, Alice is responsible for seeing to it that John submits his and vice versa. This distinction is exactly what is required in a model of group-directed imperatives, and this formulation of the imperative directed toward a distributive group illustrates the essence of our view of joint responsibility.<sup>3</sup>

We now move to explore other forms of group-directed imperatives, starting with the variant of the S modality that refers to a group of agents as a *collective*. For an imperative directed to a collective group to be satisfied, at least

<sup>3</sup> Once the imperative has been issued and the responsibility established, or the responsibility established in some other way, the means by which agents coordinate to live up to that responsibility is an important and complex topic [5,25], but not one that is within the scope of this paper.

one of the group members (possibly all of them) must be responsible for its achievement. A first attempt at a definition of this group-directed imperative may, mirroring Definition 6, be to equate it with the disjunction of individual imperatives for each member of the group, thus:

**Definition 8** ( $S_{\check{X}}A$ ) *At least one agent (in the set  $X$  being addressed), be responsible for the achievement of the state of affairs  $A$ .*

$$S_{\check{X}}A \stackrel{\text{def}}{=} \bigvee_{x \in X} S_x A$$

Suppose that Alice and John are instructed that at least one of them should prepare a presentation on next year's financial plan for a meeting tomorrow. (In this example, although not in general, it would be sensible for only one of them to see to it that the presentation is prepared.) Following the initial definition above, this imperative, directed to a collective group, would expand (where  $P$  indicates the state of affairs in which the presentation is prepared) to  $S_{\text{Alice}}P \vee S_{\text{John}}P$ . Now, if John does not contribute to the presentation, can he be held to account for not satisfying the imperative as issued? Unlike the imperative  $S_{\check{X}}A$ , we cannot say that John has failed to whole-heartedly satisfy the imperative. If Alice prepares the presentation, then the imperative is satisfied and if she does not, it is not satisfied; in the former case John has done all that he needs, and in the latter he has not. The fact that there is no notion of *joint responsibility* in this definition means that it is of little utility in capturing the concept of an imperative directed to a collective group.

Mirroring the definition of  $S_{[X]}A$  above, consider the following, stronger, interpretation of the issuing of an imperative to a collective group that emphasises group responsibility. (An alternative formulation of this definition in terms of  $S_{\hat{X}}$  and  $S_{\check{X}}$  is given in square brackets.)

**Definition 9** ( $S_{\langle X \rangle}A$ ) *Everyone in the group  $X$  is responsible for at least one member of the group achieving  $A$ .*

$$S_{\langle X \rangle}A \stackrel{\text{def}}{=} \bigwedge_{x \in X} S_x \left( \bigvee_{y \in X} S_y A \right) \quad \left[ \text{or } S_{\langle X \rangle}A \stackrel{\text{def}}{=} S_{\hat{X}} S_{\check{X}} A \right]$$

Returning to our example, the imperative instructing Alice and John to have a presentation prepared for the meeting tomorrow ( $S_{\{\text{Alice}, \text{John}\}}P$ ) will expand to the following:  $S_{\text{Alice}}(S_{\text{Alice}}P \vee S_{\text{John}}P) \wedge S_{\text{John}}(S_{\text{Alice}}P \vee S_{\text{John}}P)$ . Alice is responsible for either herself or John (or both) ensuring that the presentation is prepared and John is similarly responsible. For either Alice or John to whole-heartedly satisfy this group-directed imperative, they must take into account the activity of the other. This, therefore, enforces cooperation in the satisfaction of the imperative issued.

It is worth noting that this interpretation of an imperative issued to a collective group concurs with that defined by Rescher [43, p. 59]. Using the example of a group of students being instructed to close the door, Rescher considers the following alternative formulation: “do not let it occur that no one in the group [...] closes the door” addressed distributively to the group. Consider this alternative formulation of the example considered here: “do not let it occur that neither Alice nor John are responsible for the presentation being prepared”. This can be expressed as the following re-writing of the expansion of  $S_{\langle\{Alice, John\}\rangle}P$ :  $S_{Alice} \neg (\neg S_{Alice}P \wedge \neg S_{John}P) \wedge S_{John} \neg (\neg S_{Alice}P \wedge \neg S_{John}P)$ . In the following, however, we will propose further refinement on the types of group-directed imperatives that may be expressed so that we can capture various constraints that the issuer of an imperative may place on the group so addressed. Before doing this we define the action-oriented analogues of the state-oriented group-directed imperatives defined so far.

The group-directed imperatives  $T_{\hat{X}}\alpha$ ,  $T_{[X]}\alpha$  and  $T_{\langle X \rangle}\alpha$  are similar to their equivalents for the  $S$  modality, and are defined as follows:

**Definition 10** ( $T_{\hat{X}}\alpha$ ) *All agents (in the set  $X \subseteq \mathcal{X}$ ) are responsible for the performance of action  $\alpha$ .*

$$T_{\hat{X}}\alpha \stackrel{\text{def}}{=} \bigwedge_{x \in X} T_x\alpha$$

**Definition 11** ( $T_{[X]}\alpha$ ) *Each member of the group  $X$  is responsible for each member being responsible for the performance of action  $\alpha$ . (An alternative formulation of this definition in terms of  $S_{\hat{X}}$  and  $T_{\hat{X}}$  is given in square brackets.)*

$$T_{[X]}A \stackrel{\text{def}}{=} \bigwedge_{x \in X} S_x \left( \bigwedge_{y \in X} T_y\alpha \right) \quad \left[ \text{or } T_{[X]}A \stackrel{\text{def}}{=} S_{\hat{X}}T_{\hat{X}}A \right]$$

**Definition 12** ( $T_{\langle X \rangle}\alpha$ ) *Everyone in the group  $X$  is responsible for at least one member of the group ensuring that  $\alpha$  is done.*

$$T_{\langle X \rangle}A \stackrel{\text{def}}{=} \bigwedge_{x \in X} S_x \left( \bigvee_{y \in X} T_y\alpha \right)$$

We are now in a position to provide a generalised definition of group-directed imperatival utterances with respect to responsibility for the achievement of states of affairs and with respect to the performance of acts. In doing so, we further refine the notion by introducing the concept of a “minimal acceptable team” with respect to the group activity. By this we mean that minimum team size constraints can be placed on the group activity. For example, we can express the imperative issued to a group of  $n$  individuals that they form a circle comprising of at least 5 individuals.

**Definition 13** ( $S_{(X,n)}A$ ) *Everyone in the group  $X$  is responsible for there being established a team comprising of at least  $n$  individuals, where  $0 < n \leq |X|$ , such that that team is responsible for the achievement of  $A$ .*

$$S_{(X,n)}A \stackrel{\text{def}}{=} \bigwedge_{x \in X} S_x \left( \bigvee_{Y \in 2^X, |Y|=n} \left( \bigwedge_{y \in Y} S_y A \right) \right) \quad \text{where } 0 < n \leq |X|$$

The following special cases then follow from definition 13:

- $S_{(X,|X|)}A \equiv S_{[X]}A$ . In this case, the only permissible team is that which consists of all those addressed.
- $S_{(X,1)}A \equiv S_{\langle X \rangle}A$ . In this case, the team may contain one or more individual; i.e. the imperative is satisfied if all those addressed ensure that at least one individual in the group (possibly all of them) is responsible for  $A$ .
- $S_{(\{x\},1)}A \equiv S_x S_x A$  and by TS,  $S_x S_x A \rightarrow S_x A$ .

Thus, this definition provides a general definition of individual- and group-directed imperatives with the added advantage that the issuer of the imperative can place minimum cardinality constraints upon the group required to achieve the goal indicated.

An equivalent definition for individual- and group-directed imperatives with respect to an action to be performed may be constructed.

**Definition 14** ( $T_{(X,n)}\alpha$ ) *Everyone in the group  $X$  is responsible for there being established a team comprising of at least  $n$  individuals, where  $0 < n \leq |X|$ , such that that team is responsible for the achievement of  $\alpha$ .*

$$T_{(X,n)}\alpha \stackrel{\text{def}}{=} \bigwedge_{x \in X} S_x \left( \bigvee_{Y \in 2^X, |Y|=n} \left( \bigwedge_{y \in Y} T_y \alpha \right) \right) \quad \text{where } 0 < n \leq |X|$$

It is worth illustrating these definitions with a concrete example. Suppose that two or more members of a group of three,  $\{x, y, z\}$ , should attend a meeting (action  $\mu$ ). This can be expressed as  $T_{(\{x,y,z\},2)}\mu$ , and expands to:

$$\begin{aligned} & S_x ((T_x \mu \wedge T_y \mu) \vee (T_x \mu \wedge T_z \mu) \vee (T_y \mu \wedge T_z \mu)) \wedge \\ & S_y ((T_x \mu \wedge T_y \mu) \vee (T_x \mu \wedge T_z \mu) \vee (T_y \mu \wedge T_z \mu)) \wedge \\ & S_z ((T_x \mu \wedge T_y \mu) \vee (T_x \mu \wedge T_z \mu) \vee (T_y \mu \wedge T_z \mu)) \end{aligned}$$

Each line (i.e. each outer conjunct) expresses, loosely, each agent's joint responsibility to the minimum acceptable team. The different configurations of that minimum acceptable team are then expressed by the disjuncts. Of course,

all three may attend the meeting, which simply leaves multiple configurations met.

We might imagine a situation in which  $x$  and  $z$  agree to attend the meeting. If  $z$  falls sick in the interim, and cannot attend, the second and third disjuncts of each conjunct are assuredly false, and  $x$  and  $y$  might reason normatively that the first disjunct must be fulfilled, i.e. that they are the two that must attend. This brief and informal scenario offers a first example of how an agent might acquire responsibility: an issue to which the next section turns in more detail.

## 4 Acquiring Responsibility

### 4.1 Acquiring Responsibility through Action

The simplest way for an agent to be responsible for something is through direct action. That is, if an agent carries out an action then, *ceteris paribus*, that agent is responsible for the execution of that action. Perhaps the simplest way to capture this is to construct an axiomatic representation that links successful action to responsibility for that action:

**Definition 15 (RNR)**

$$\frac{\alpha^x}{\top_x \alpha^x}$$

At first glance, this appears to be re-introducing necessitation, RN, by the back-door, after having worked so hard to exclude it in the development of the syntax and semantics, in order to preserve Hamblin’s intuitions. But the resemblance is superficial. RN captures the relationship between a proposition, unrelativised, and relativised responsibility. What Jones and Sergot found so galling — and what Chellas found so intuitive — was that logical truths should be the responsibility of any agent whatsoever. What (RNR) is capturing, however, is a much more limited notion of what Jones and Sergot referred to as building a “logic of successful action” — namely, that a *specific* agent’s action should imply responsibility for that action by that agent. Although it has formal similarity to RN, the relativised version RNR is a much weaker notion.

Even this, however, is a little too strong. There are cases where agents of direct actions may, arguably, be absolved of responsibility for their actions. This is perhaps clearest in the legal notion of *diminished responsibility*. There are examples in some legal systems of cases in which defendants have argued that their actions were not voluntary, or were carried out when not of sound



mind, and that therefore culpability is reduced or eliminated. This caveat suggests that the rule expressed by RNR needs to be weakened. A defeasible or default logic would be a good candidate for capturing the intuition that usually, committing an act incurs responsibility — but that exceptions may be conceivable. This might be represented in the usual default style as,

**Definition 16 (DefRNR)**

$$\frac{\alpha^x : \text{T}_x \alpha^x}{\text{T}_x \alpha^x}$$

It is perhaps tempting to try to push further in this direction, and in some way associate execution of actions with responsibility for their consequences. Under certain conditions, one might want to try and enforce that

$$\text{T}_x \alpha^x \rightarrow \text{FA}$$

(i.e. that responsibility for the execution of a given action  $\alpha$  has some effect  $A$  that holds at some point later), and thence that

$$\text{if } \vdash_M \text{T}_x \alpha^x \rightarrow \text{FA} \text{ and } \vdash_M \text{T}_x \alpha^x \text{ then } \vdash_M \neg \text{S}_x \neg \text{FA}$$

Such a relationship appears to be close to Hamblin’s notion of an agent’s partial *i*-strategy for ensuring that the content of some imperative is not prohibited. The attempt, however, is wrong-headed. What these relationships are defining is a *logical* relationship between actions and their consequences. Though there is obviously some relationship, Hamblin is at pains to point out that there is an important gap between, as he puts it, logically possible worlds and physically possible worlds. Enshrining the link between actions and their effects in a logical, axiomatic relationship is, for Hamblin, too strong. It also risks conflating the very distinction he constructed at the centre of the theory. As a result, we here leave the link between S-formulae and T-formulae undefined, leaving it to a practical reasoning component to implement an appropriate model of the link (i.e. to implement an appropriate model of causality). One important advantage of this approach is in considering the ethics of delegation; an issue to which we return in Section 5. Given the simplicity of the S and T language it is relatively straightforward to map from a number of traditional approaches to the sort of reasoning required. At one end, there are purely logical approaches that seek to construct a model of epistemically embedded rationality, in which intentions drive the generation of goals that in turn drive (or prune) action selection (such as [48]). At the other, are engineering-oriented approaches that aim to construct and modify plans based on causal reasoning (such as POCL planners that trace their ancestry back to UCPOP [38]). Both these classes of approach have clear representation of actions or states (or, in a limited sense, both), and assume that the “raw ma-

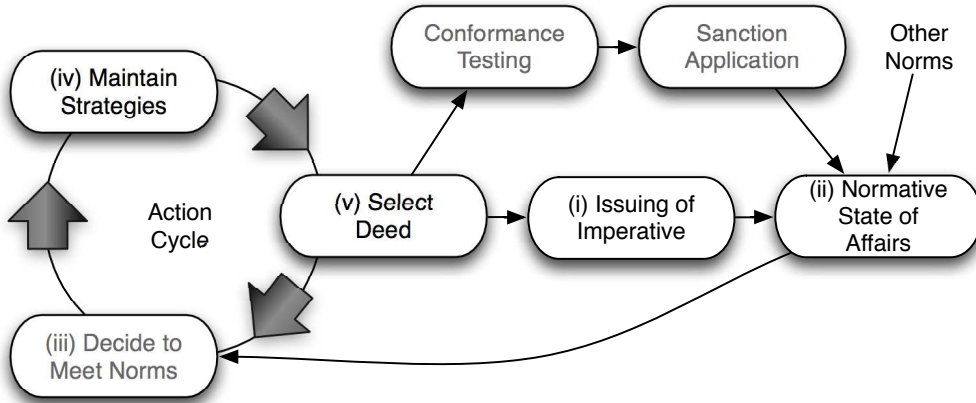


Fig. 4. The imperative-normative-action cycle

terials” of descriptions of those states or actions are available. It is those raw materials that are described here.

It may be that the richness of Hamblin’s underlying model – and possibly its formalisation in [42] that is employed here – are sufficient to support a detailed analysis of the interactions between actions, their effects, and foreknowledge of those effects. Hamblin’s description of  $W_{phys}$  and its position with respect to logical and temporal possibility suggests that this may be so. But such an account is beyond the scope of the current paper and is not a necessary prerequisite for the construction of a logic of delegation.

#### 4.2 *Acquiring Responsibility through Delegation*

Hamblin’s original motivation for developing action state semantics was to be able to explain imperatives and construct a logic that was designed to cope with them. We are now at a point to return to that motivation. The aim here, broadly, is to explain the five-step process (Fig. 4) from (i) the issuing of an imperative, to (ii) a deontic state in which requirements of responsibility are assigned, then (iii) an agent’s whole-hearted satisfaction of that imperative, which (iv) leads to the agent’s responsibility for the state or action thereby (v) discharging of the deontic demand.

Broadly, an agent acting within a normative environment follows the pattern shown in Fig. 4. The action cycle comprises an agent recognising and selecting one or more norms to meet, and then, in Hamblinian terms, maintaining partial strategies appropriate to the satisfaction of those norms. Selection of deeds is then constrained by those strategies. Communication is one specific type of action, and uttering an imperative is one specific type of communication. Issuing an imperative can update the normative state to introduce some

new norm, which is often one that affects some other agent. The selection of deeds can be compared to some standard and conformance testing applied, resulting in sanctions being applied if conformance testing fails. Here, we are interested in the central path from an agent's strategies to the deeds it selects that are consistent with those strategies, and in particular, communicative deeds involving imperatives, which in turn update norms. A key notion in understanding this path is Hamblin's concept of whole-hearted satisfaction.

#### 4.2.1 Whole-hearted Satisfaction

Wholehearted satisfaction is based upon the notion of a strategy. A strategy for a particular agent is the assignment of a deed to each time point. A partial  $i$ -strategy is then a set of incompletely specified strategies, all of which involve worlds in which  $i$  is extensionally satisfied. The wholehearted satisfaction of an imperative  $i$  by an agent  $x$  is then defined as being  $x$ 's adoption of a partial strategy and the execution of a deed from that strategy at every time point after the imperative is issued.

A Hamblinian world  $w \in W$  is defined such that for every time point in  $T$  there is:

- (1) a state from the set of states  $S$ ,
- (2) a member of the set  $H$  of 'big happenings' (each of which collects together all happenings from one state to the next), and
- (3) a deed (in  $D$ ) for every agent (in  $X$ ), i.e. an element from  $D^X$ .

The set  $W$  of worlds is, therefore, defined as  $(S \times H \times D^X)^T$ . The states, happenings and deed-agent assignments of a given world  $w$  are given by  $S(w)$ ,  $H(w)$  and  $D(w)$ .

The next step is to let  $j_t$  be a history of a world up to time  $t$ , including all states, deeds and happenings of the world up to  $t$ . Thus  $j_t$  is equivalent to a set of worlds which have a common history up to (at least) time  $t$ .  $J_t$  is then the set of all possible histories up to  $t$ ; i.e. all the ways in which the universe could have evolved up to time  $t$ . A strategy  $q_t$  is then an allocation of a deed to each  $j_{t'} \in J_{t'}$  for every  $t' \geq t$ .<sup>4</sup>  $Q_t$  then denotes the set of all possible strategies at time  $t$ .

Let the possible worlds in which the deeds of agent  $x$  are those specified by strategy  $q_t$  be  $W_{\text{strat}}(x, q_t)$ , and the worlds in which an imperative,  $i$ , is extensionally satisfied be  $W_i$ . A strategy for the satisfaction of an imperative  $i$  (i.e. an  $i$ -strategy) can, therefore, be defined as follows: A strategy  $q_t \in$

---

<sup>4</sup> This notion of a strategy has an intensional component, since it prescribes over a set of possible  $w$ , rather than picking out, at this stage, the actual world.

$Q_t$  is an  $i$ -strategy for agent  $x$  if and only if the worlds in which  $x$  does the deeds specified by  $q_t$  are also worlds in which  $i$  is extensionally satisfied:  $W_{\text{strat}}(x, q_t) \subseteq W_i$ .

In practice, however, it is not feasible for an agent to select a particular strategy in  $Q_t$  at time  $t$  that specifies every deed for every time  $t'$  after  $t$ . For this reason, an agent will adopt a *partial*  $i$ -strategy. A partial  $i$ -strategy is a disjunction of  $i$ -strategies,  $Q'_t \subseteq Q_t$ , and the set of worlds in which  $x$  adopts this partial  $i$ -strategy is  $W_{\text{strat}}(x, Q'_t)$ .

With this grounding, the wholehearted satisfaction of an imperative,  $i$ , can now be defined. An agent  $x$  may be said to wholeheartedly satisfy an imperative  $i$  issued at  $t$  if and only if for every  $t' \geq t$ :

- (1)  $x$  has a partial  $i$ -strategy,  $Q'_{t'}$ ; and
- (2)  $x$  does a deed from the set of deeds specified by that  $Q'_{t'}$ .

For further details, the reader is referred, of course, to Hamblin's original monograph [23], and also to [53], which provides more detail on the role of such a model in the wider context of dialogue and, in its appendix, a more complete set-theoretic précis of Hamblin's model.

#### 4.2.2 Responsibility from Whole-Heartedly Satisfying

There are at least two ways in which an agent might be said to be responsible for something. The first, as we have seen in Section 4.1, is by direct execution. If an agent,  $x$ , is the direct executor of some action,  $\alpha$  (i.e.  $\alpha^x \in D$ ), then, other things being equal, that agent can be said to be responsible for that action. Responsibility, however, is a broader notion that is well matched by Hamblin's construction of whole-hearted satisfaction. The same sense of necessary involvement is pivotal, or, as Hamblin puts it, "[Something cannot count] as wholeheartedly satisfied if it is possible to say of it, *He wouldn't have done it if it hadn't been for so-and-so*, or, *It only came about by accident*, or *It would have come about anyway, what he did was irrelevant to it (or impeded it)*. Conversely, even when extensional satisfaction is lacking we sometimes want to say, *Yes, but it wasn't his fault*, or *He did everything he could*." [23, p155].

The whole-hearted satisfaction of some normative state of affairs (that was brought about by an imperative, or possibly some other societal or contextual process) thus represents a more general picture of agentive responsibility, of which DefRNR is a specific instance. One way for an agent to whole-heartedly satisfy some expression is by building and maintaining a strategy in which, at an appropriate moment, the agent is a direct executor of an action that corresponds to the expression. Even such postponed involvement may be un-

necessary, however. An agent may be able to be responsible for something simply by establishing an appropriate normative state — specifically, by issuing an imperative. In the context of an imperative, the first link between the ST-notion of responsibility and Hamblin’s wholehearted satisfaction starts to emerge. The T-formula, for example, that captures the statement of responsibility of a recipient of an imperative can be said to *constitute* wholehearted satisfaction of that imperative. Having statements in the logic ST constitute statements of whole-hearted satisfaction in this way is a crucial step in understanding how delegation can effect the transfer of responsibility via imperitival utterance. It is to a more formal understanding of this relationship that we now turn.

#### 4.2.3 Delegation by Imperitival Utterance

Given an appropriate social context, one agent may alter the normative state in some way for another agent. One mechanism for effecting such a change is through an act that constitutes imperitival communication. To demonstrate how such communication might itself constitute whole-hearted satisfaction of some other norm, we use a small example.

Consider an agent  $x$  that is obliged to submit a report, action  $\alpha$ , as a requirement of the position it holds in the organisation. As a consequence of this requirement,  $x$  forms an intention that  $\alpha$  should be carried out. Let us further assume that  $x$ ’s reasoning mechanisms determine that proactivity is appropriate; i.e., that it should be responsible for  $\alpha$  being carried out,  $\text{T}_x\alpha$ . Let us imagine that  $x$  has no resources available for performing  $\alpha$  itself, but that there is another agent,  $y$ , over whom it has authority. In transforming its intention into a partial strategy, agent  $x$  might issue to  $y$  an imperative  $i$  expressing that  $y$  should ensure that  $\alpha$  is done. We use an abstract form for this imperative, where the predicate  $!$  is instantiated by a request or a command or some indirect speech act, to convey responsibility for  $\alpha$ ; we continue to use the superscript convention to express agentive execution, so that  $!^x$  is an imperative issued by  $x$ . The addressee of the imperative can be marked as a subscript,  $!^x_y$ , but to keep things syntactically simple, we adopt the convention that if not marked explicitly, the addressee is taken to be the agent (or group) to which the S or T modality is relativised. Thus the imperative comes out as  $i = \text{T}_y\alpha!^x$ . Given the appropriate context (the existence and mutual knowledge of the authority relation, for example), this utterance creates a new normative expression such as  $\text{SCOMM}(x, y, \text{T}_y\alpha)$ . (We use the notation of [8] because it is simple and intuitive, but any appropriate language might be substituted.) This normative state in turn influences agent  $y$  which develops (or, more accurately, is obliged to develop) a strategy for fulfilling  $i$ , namely, a partial  $i$ -strategy.

There may be any number of ways for  $y$  to whole-heartedly satisfy  $i$ , i.e. there may be a diverse set of partial  $i$ -strategies predicating different actions for  $y$ . The most straightforward is a strategy that requires  $y$  to perform  $\alpha$  directly. At that point (i.e. in that world),  $\alpha^y$  becomes true, and, by DefRNR, thence  $\top_y \alpha^y$ . Clearly,  $\alpha^y$  provides extensional satisfaction of the imperative  $i$ , but it is  $\top_y \alpha^y$  that constitutes the whole-hearted satisfaction of  $i$  — we might mark this syntactically as  $whs(y, i)$ , or, more explicitly,  $whs(y, \top_y \alpha^y)$ .

What we need in order to close the loop is a relationship between whole-hearted satisfaction and delegated responsibility, which we are now in a position to define axiologically, with the rule of inference for satisfaction, RS:

**Definition 17 (RS)**

$$\frac{whs(y, \top_y \alpha^y)}{S_x \top_y \alpha}$$

In this way, RS delivers a statement of responsibility that corresponds to the whole-hearted satisfaction of an imperative. The result of this inference then yields two conclusions. The first, unsurprisingly, is the responsibility of the subordinate agent,  $y$ , because by T,  $S_x \top_y \alpha$  gives  $\top_y \alpha$ . Of course, in the example we are looking at here,  $\top_y \alpha$  is inferrable from  $\top_y \alpha^y$  by generalisation [42]. But in the general case in which whole-hearted satisfaction has been achieved by means other than direct execution,  $\top_y \alpha$  may not hold — so it is encouraging, therefore, that it is inferrable by other means.

More importantly,  $S_x \top_y \alpha$  also gives, by our axiom of delegation QT, that  $\top_x \alpha$ . In other words,  $x$ 's responsibility for  $y$ 's responsibility for  $\alpha$  being executed implies  $x$ 's responsibility for  $\alpha$  simpliciter. This is vital, because it provides the mechanism by which  $x$  can reason (for example, by planning or backchaining) that issuing the imperative  $\top_y \alpha^y$  will serve its intention of being responsible for  $\alpha$  (i.e. of  $\top_x \alpha$ ).

The route from  $x$ 's intention  $\top_x \alpha$ , generating the imperative  $\top_y \alpha^y$ , causing the social commitment  $SCOMM(x, y, \top_y \alpha)$ , which in turn obliges a combination of the extensional satisfaction of  $\alpha$  and a partial  $i$ -strategy (such as the one by which  $\alpha^y$ , and therefore  $\top_y \alpha^y$ ) that delivers  $whs(y, \top_y \alpha^y)$ , which in turn implies  $S_x \top_y \alpha$  yielding fulfilment of the original intention is the model of delegation supported by the logic of ST, and is summarised in Fig. 5 (and the path through the partial  $i$ -strategy labelled (1)).

### 4.3 Multi-Step Delegation

Section 4.2.3 described one example of a simple partial  $i$ -strategy that an agent might adopt in meeting the demands of whole-hearted satisfaction. There are,

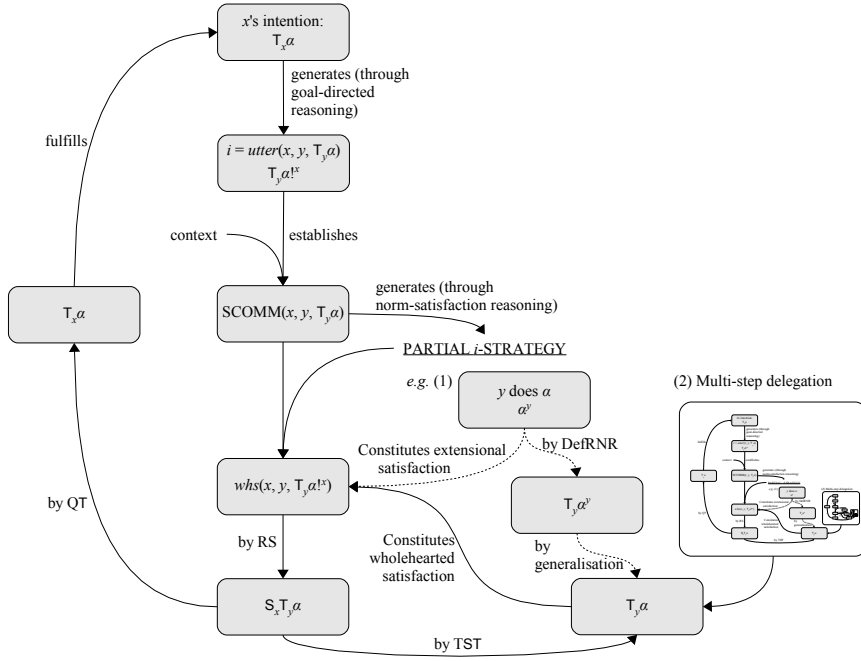


Fig. 5. A logical account of delegation

of course, any number of such strategies that an agent might adopt in such a situation. One interesting alternative is where the agent decides to delegate the task further. So, in our earlier example,  $y$  may decide that submitting the report is best done by a subordinate, agent  $z$ , to whom  $y$  must delegate the task.

Thus, the social commitment  $SCOMM(x, y, T_y\alpha)$  in Fig. 5 can be seen in the second partial  $i$ -strategy to generate the intention in  $y$  that  $T_y\alpha$ . That intention in turn generates the imperative  $T_z\alpha!$ , causing the social commitment  $SCOMM(y, z, T_z\alpha)$ , which in turn obliges a combination of the extensional satisfaction of  $\alpha$  and a partial  $i$ -strategy for  $z$ , such as the one by which  $z$  performs  $\alpha$ , i.e.  $\alpha^z$  and therefore  $T_z\alpha^z$ ) that delivers  $whs(z, T_z\alpha!$ , which in turn implies  $S_y T_z\alpha$  yielding fulfilment of  $y$ 's original intention. In other words, the entirety of Fig. 5 can be embedded as the partial  $i$ -strategy for  $y$  (with  $z$  substituting for  $y$ , and  $y$  for  $x$ ).

This embedding of one delegative step within another is exactly what one would hope for, since multi-step delegation of this kind is a very natural activity. There are, however, cases in which delegation of this sort is undesirable. Let us modify the example from Section 4.2.3 slightly, and imagine that agent  $x$  delegates the work on the report to  $y$  because  $y$  is the best report writer in the team. In such a situation,  $x$  explicitly does not want  $y$  to delegate the

task further: how can  $x$  effect such delegation? The answer lies in considering a “guard” condition that is added to the imperative: not only must the addressee be responsible for the action, but, furthermore, they must not be responsible for further delegation. There are strong echoes of Lindahl [30] and Hart [24] here in constructing the perimeter of rights using guards in this way. The challenge is that, as we have seen, there is no identifiable “delegate” action that can be prohibited (and nor would we want there to be such an action; it seems self-evident to us that it is important to allow delegation to be achieved through many if not all of the existing primitives or mechanisms of an existing communication language). Nor can we circumvent the absence of a delegate action by trying to identify a unique part of the state that can be identified with the postcondition of delegation (apart from anything else, this would conflate the action/state distinction introduced by Hamblin and preserved in the language of ST). A final complexity lies in the fact that just because one agent  $x$  is prohibited from delegating an action  $\alpha$  to some other agent  $y$ , there is no reason to assume that  $y$  might not already have received imperatives, or otherwise be socially committed to various actions — including, perhaps,  $\alpha$ . The prohibition on  $x$ ’s freedom cannot, therefore, be expressed solely in terms of the (potential) commitments of the agents to whom it might (potentially) delegate.

The solution turns (again) upon Hamblin’s notion of whole-hearted satisfaction, and its link to responsibility. By expressing the guard as itself an expression of the addressee’s responsibility, whole-hearted satisfaction does not preclude extensional satisfaction by other means. Thus, for our running example, if  $x$  wishes to delegate  $\alpha$  to  $y$  in such a way that  $y$  does not delegate further, the appropriate imperative is  $(T_y\alpha \wedge \forall z \neg S_y T_z\alpha)!^x$ . In this way,  $y$  is directed to take on responsibility for  $\alpha$ , but in addition (in the guard) is required not to be responsible for the state in which any other agent is responsible for  $\alpha$ . Of course some other agent may, in fact, wind up with such responsibility (or may have it already), but so long as that responsibility has not itself been brought about by  $y$  (i.e. is not  $y$ ’s responsibility), then the compound imperative can still be whole-heartedly satisfied.

Finally, the reverse situation is also possible:  $x$  may wish to direct  $y$  as to execution of  $\alpha$ , for instance by demanding that it be delegated further. Examples of explicit multi-step delegation imperatives typically involve some jump in a chain of command, from superior to inferior — e.g., “No, I’m sorry, this won’t do. I need you to find someone who can do the plastering *properly*.” Again, no delegate action is required in order to capture this, merely an expression such as  $\exists z S_y T_z\alpha!^x$  (where  $T_y\alpha$  is inferable from  $S_y T_z\alpha$  by QT). Of course, such interference with the addressee’s fulfilment of the imperative is not limited to demands for further delegation: the speaker could indicate means, timescale, or groups, amongst guard conditions; the speaker could also, using the language of Section 3.2 demand not just performance or assurance of actions



or establishment of states, but also forbearance, interdiction or prohibition. We emphasise the delegate/no-delegate guards in particular to lay a foundation upon which future work can build imperative-based characterisations of Lindahl- and Hart-style contractual relationships.

This concludes our presentation of a logic of delegation that answers many of the most challenging questions in building a comprehensive model of the concept. The issues surrounding the locus of responsibility for delegated activities, the Refref conjecture and the complexities of group-directed imperatives have been addressed, characterisations of these concepts have been offered, and mechanisms whereby responsibility may be acquired have been presented. In the following section we discuss some important related research, attempting to capture the breadth of work in this important area. Our conclusions follow in Section 6.

## 5 Related Work

Comparison with Belnap’s conception of Refref has been explored in detail in Section 3.2, and differences between the underlying language and that of the *stit*s in [42]. Belnap *et al.*’s [4] rich and detailed theory also explores imperatives and the links between *stit*, communication and deontic states (although does not explore delegation explicitly). One interesting observation by way of comparison with part of this work is their discussion [4, Ch. 12] of Marcus’s [32] “unpretentious” example, *Parking on highways ought to be forbidden*. For Belnap *et al.* the interpretation of this example is:

*Oblg*: $[\Gamma \text{ dstit}: \exists\beta[(\beta \in \Gamma) \ \& \ [\beta \text{ dstit}: \text{Sett}: \text{Will-always}:\forall\alpha\forall x \text{ Frbn-if-can-do}:[\alpha \text{ dstit}: P\alpha x]]]]]$

The first part of this identifies an individual actor,  $\beta$ , from a particular group,  $\Gamma$  (“*the authorities*” or “*they*”), and the next part the temporally and spatially quantified prohibition (all agents  $\alpha$  forbidden from parking  $P$  on a highway  $x$ ). The theory presented here has little or nothing to say about the deontic aspect of this example. It does, however, allow us to describe the appropriate states, i.e. to focus not upon the ideality but instead upon the actuality (of the authorities’ responsibility for everyone’s responsibility for not parking on highways). By making use of both collective and distributive groups, and the distinction between state and event-oriented responsibility, we have:  $S_{\langle\Gamma\rangle}G\forall xT_{[A]}\neg Px$  (where the set  $A$  is taken to be the domain of  $\alpha$ , which is left implicit in Belnap *et al.*’s interpretation).

The work of Governatori and colleagues is also worthy of mention here in relation to the logical underpinnings of our model of delegation. Governatori

and Rotolo [21], for example, elegantly fix Elgesem’s [19] account of action which in its Success and Non-accidence conditions is similar to the approach presented here. In Avoidability, however, Elgesem, and Governatori and Rotolo diverge from the RT-type logic that gives our account of delegation its rich notion of responsibility.

We have discussed elsewhere [42] how the underlying formal model compares with that of Singh [44,45] whose programme of research represents one of the most significant, sustained investigations of the area in AI. Although Singh [46] also takes Hamblin as inspiration and his starting point, he does not address delegation directly in the context of his work with WSAT. Delegation does appear explicitly in his more engineering-oriented work [47], where the types of delegation in which we are interested appear explicitly as structural patterns. To suit the target audience for that work, they are not tied to underlying formal explication and are highly simplified (being reduced to atomic actions). Finally, Singh also comments (personal communication) that he does not like the term “imperative” because it is based on natural language syntax rather than semantics, and thereby conflates the diverse semantics of directives, permissives and prohibitives. For the work here we are keen to disentangle the semantics of a given communicative utterance from the subsequent normative state, so that very diversity of semantics aids our purpose. (It is also convenient to be able to retain the term for its intuitive simplicity and to retain the explicit link to Hamblin’s original work.)

Further discussion of research related to our underlying logic is given in [42], here, however, we focus more on related work within multi-agent systems.

Pacheco and Santos [37] formulate the issue of delegation in the context of role-based organisations, and, like us, rely upon a deontic characterisation of state to contribute to an understanding of how delegation is achieved. Unlike the model presented here, however, Pacheco and Santos place heavy restrictions on delegation, including, for example, requiring that the delegator “also transfers [...] all the resources required”, presupposing both that those resources exist, that the delegator is aware of them, and has the power to require the delegatee to employ them. This, of course, arises ultimately from conflation of delegation of state (which can leave methods unspecified) and delegation of action (which does not). Pacheco and Santos also model joint agency only in respect of the delegator (which is achieved through a simple conjunction over contributing agents): the more complex delegation to multiple agents would, for Pacheco and Santos, require formation of a institutional agent and some separate system for distributing responsibility in that institution. We argue, and have shown, that the process of delegation is intimately tied to individual and joint responsibility, and that a definition of either is incomplete without reference to the other. Finally, Pacheco and Santos, reasonably enough, limit themselves to a deontic characterisation of delegation in isolation from

features such as motivation, reasoning and communication. We have demonstrated that by using as a basis a rich account such as Hamblin's that includes communicative and social aspects, we can construct a model of the entire cycle to explore what happens, for example, in cases of multiple-step delegation.

Also within the context of agent reasoning mechanisms, Boella and van der Torre [6] describe the "social delegation cycle" which can be seen as a generalised version of the imperative-normative-action cycle presented in Fig. 4, and though both [6] and [7] are rather limited in their treatment of delegation as transfer of responsibility, they provide a rich multi-modal logic that provides a context in which much of the work in the current paper could be set. Van der Hoek and Wooldridge [50] offer a mechanistic account of delegation that uses dynamic logic to describe how the control of propositional variables can be passed between agents. That control is equated with power (i.e. power to alter the Boolean value of the variable), and transfer of control with delegation. Although this approach provides an interesting abstract model and may be a way to implement some kinds of simulations of delegation (through work such as the Logic Programming-oriented [29]), it fails to handle a number of issues that are central to the model presented here, *viz.*, delegation to groups, delegation with unspecified means, responsibility across delegation, the interactions between communication and delegation, and so on.

Kumar *et al.* [28] make the distinction between "a group doing an action as an entity (or meta-agent)" and "everybody in a list of individuals performing an action". This distinction is possibly due to the formulation of their generalised request action. In specifying the generalised request action, Kumar *et al.* focus on the distinction between agents to whom a request is directed and those who simply overhear it. This permits the definition of a request using Cohen and Levesque's mentalistic logic [12] such that the issuer does not know who the intended actor is within the group that the request is directed towards. A request to a group treated as a "meta-agent" can then be defined, but the model is also able to capture requests directed to individuals as before. There are a number of limitations of this approach including the underlying logic, which relies on pseudo-states that express that specific acts have been done (see Reed and Norman [42] for more discussion on this issue). With respect to capturing the concept of group-directed imperatives, however, Kumar *et al.* [28] go little further than Rescher [43] who presents a thorough discussion of the importance of distinguishing between imperatives directed towards a collective or distributive group and a practical model of these concepts. Kumar *et al.* do, however, point towards how agent communication language specifications grounded on belief-desire-intention logics can be extended to refer to groups of agents. A more significant contribution in this regard, although only capturing the distributive case, is the model of establishing collective intention through dialogue proposed by Dignum *et al.* [17]. This model represents one of the first attempts to tie together a form of dialogue (in their case a form

of persuasion dialogue based on RPD [53]) to resultant mental states of the participants (in their case an intention that is common to all agents within the group engaging in the dialogue). Davis and Morgenstern [15] tackle the link between communication and planning in a multi-agent setting in a formal but pragmatic manner. Rather than addressing the issue of how one agent persuades another to adopt an intention, they assume agents are cooperative and pre-allocate time slots to others in anticipation of requests for plans to be executed. David and Morgenstern’s contribution is in formally capturing some of the more simple forms of delegation, including group-directed delegation, considered in this paper and in the links with a model of knowledge. One of the examples involves an agent issuing a request to a group of agents, one of which accedes to the request. The question of the group collectively deciding which agent accedes to the request is, however, not considered; in the example used by Davis and Morgenstern there is a unique resource that is requested, and hence the only agent to respond is the one that has that unique resource; i.e. the question of how a group coordinates to respond to a request is not considered. By bridging the gap between requests and planning, however, Davis and Morgenstern [15] do offer a possible realisation of the notion of whole-hearted satisfaction: the responsible agent accedes to requests by adding actions in their plans (during one or more of the time slots pre-allocated to the requesting agent) and ensuring that no other action interferes with the satisfaction of this request.

One issue that Rescher [43] points out, is that not only is it possible that the recipient of an imperative (or any other communicative act) be a group, but the issuer (or source) of that act may be a group (collective or distributive). Rescher [43] uses a number of examples to illustrate the various possibilities. These include: “*Group (Collective) to Group (Collective)* Court order to a corporation to divest itself of certain holdings (in violation of antitrust statutes)” [43, p. 13]. Jones and Sergot [26] use similar examples to illustrate their “counts as” connective; for example, “*x*’s uttering the words ‘I pronounce you man and wife’ counts (in [society] *s*) as a means of guaranteeing that *s* sees to it that [the recipients of the declaration] are married”. Jones and Sergot do not, however, confine their theory to communicative acts, but present a general theory of agents acting on behalf of a group. An analysis of the utterance of an imperative by an individual on behalf of a group may, therefore, be related to Jones and Sergot’s [26] notion of “counting as in a society”. This is a necessary element of a complete theory of delegation, and is an important avenue for future research, some first steps in which have been presented in [51].

On commands in dialogue and the context of the issuing of a command, Rescher [43] states that, “[g]enerally speaking, the source should have some entitlement or authority for giving a command to its recipient”. This means that a command (or any imperative) could be questioned by its recipient regarding the authority of the source and the grounds for it being issued. Understanding

how the issuing of an imperative fits into the wider structure of inter-agent dialogue may influence the design of flexible agent communication protocols. Recently this issue has been addressed by Atkinson *et al.* [1]. Building upon the earlier work by Girle [20], Atkinson *et al.* present argumentation schemes [54] for the issuing of commands within dialogues and demonstrate how these may be employed within a broader command dialogue protocol (CDP) and in the context of organisational relationships between dialogue participants. This represents an important first step to building a complete, pragmatic theory of imperatives in dialogue, which complements the model presented in this paper.

Castelfranchi and Falcone [9] pose a number of interesting questions that a theory of delegation must answer, focussing on the social (or organisational) context within which delegation takes place. These questions serve to clarify what, in their view, is required for a comprehensive theory. They address the nature of the object of delegation, the nature of the relationship between the parties involved, the autonomy of the agent to whom a task is delegated, what is meant by “on behalf of”, and the issue of trust between the parties. For Castelfranchi and Falcone, the object of delegation is a task, goal pair, capturing their intuition that it is only meaningful for delegation to refer to a specific act along with the outcome that this act is required to produce. In contrast, our model allows agents to delegate specific acts (captured by the T modality) without requiring reference to the intended outcome, or to a specific outcome (captured by the S modality) without specifying the means by which this outcome is to be achieved (an approach also advocated by Morgenstern [34]). In this respect, our model is more flexible; it is possible to tie tasks and goals intimately together by constraining the model in this way, but this is neither necessary nor essential.

Separating the object of delegation from its consequences, or from its side-effects, is also important when issues of ethics are a concern [27]. Consider, for example, a failure in an automated aircraft control system leading to loss of life in a crash (the cause of the failure having been determined by an inquiry). The plane manufacturers and maintenance engineers (as a collective group) were responsible for providing an effective and safe aircraft. Suppose that the conclusion of the inquiry was that the plane manufacturers have failed to fulfill their responsibility. To draw this conclusion, however, a domain-specific model of causality is required. It might be argued that responsibility lies with the aircraft designers if the design was at fault, but this does not take into account testing and other phases of development. Responsibility for the side-effects of actions (given a specification of the actions within a domain) has been explored by Grossi *et al.* [22]. In this complementary research an agent can be said to be responsible for causing some state of affairs  $\phi$  to hold if it just performed some action  $\alpha$  and  $\phi$  would not have necessarily been the case if  $\alpha$  was not performed by that agent. Such causal inference is essential for

a model of reasoning about agentive responsibility, in the sense of *what an agent can be held responsible for*. It is, however, not something that should be a component of a logic of delegation and agentive responsibility, in the sense of *what an agent is responsible for, following the issuing of an imperative*.

The model proposed by Castelfranchi and Falcone also fails to address the important issue of delegation to groups of agents, something that is essential for a comprehensive theory, as argued in this paper and by others. One of the important emphases of the theory developed by Castelfranchi and Falcone is the link between delegation and the mental states of the participants that are *required* (for a decision to delegate to be made) and *consequent* (i.e. resultant). As noted by Lorini *et al.* [31], this link is not something that has featured strongly in our model until now; a limitation that we have partially resolved in Section 4. There is further research to be conducted in this area, however. The link between motivations for delegating tasks and assessments of the competencies of agents to whom a task is delegated, for example, has not been fully explored in this paper (although there is a substantial relevant literature in the area of trust [41]). To give a concrete example, it may be reasonable for a teacher to delegate a task to a group of students that the teacher does not believe competent to complete. Within the research reported here, however, we abstract away from these issues to focus on the meaning of responsibility and delegation. Concepts of belief, trust, monitoring, etc. will necessarily be at the core of a complete model of actual responsibility [35], but this does not diminish the fact that, in principle, there is a transfer of responsibility albeit defeasible. Conte and Paolucci [13] further explore the social context of delegation, and explore the associated concepts of “counting upon” others for activities and of the accountability of agents. In this complementary research the authors address the concepts of shared and collective responsibility, where collective responsibility corresponds to that of the collective group and shared responsibility corresponds to that of the distributive group as used by Rescher [43] and in this paper. Conte and Paolucci provide a detailed analysis of the social context underpinning these and other related notions including power, provide some useful examples that serve to ground this analysis and relate these concepts to other related research including that of Jones and Sergot [26] discussed above.

## 6 Conclusions

We set out to develop a logic of delegation. The foundation of the approach is the formal characterisation of the modalities S and T developed in full in [42]. Using this starting point, the model presented here ties together the axiological and semantic aspects of delegation both individually and to groups, both singly and in series, both with positive and negative responsibility. Our aim

in the development of this model has been to bridge the gap between philosophically well-grounded conceptions of responsibility and practical, implementable logical systems that can support delegation. We have shown how it can both contribute to a philosophical understanding of, for example, forbearance, whilst simultaneously providing a practical account of group-oriented communication and the acquisition of responsibility.

The paper provides, for the first time, a detailed analysis of the anatomy of delegation in terms of the logical, normative and inferential aspects of an agent's world. This analysis can serve as a starting point for building models of agents that use responsibility and the transfer of responsibility as key components in their solo and social reasoning. Though the picture presented here makes a number of simplifying assumptions, it benefits from the richness and detail of Hamblin's account of imperatives, and provides those benefits in full to designers of heterogeneous agent systems.

## Acknowledgements

We are indebted to our reviewers for detailed, insightful and useful comments on an earlier draft.

## References

- [1] K. Atkinson, R. Girle, P. McBurney, and S. Parsons. Command dialogues. In I. Rahwan and P. Moraitis, editors, *Proceedings of the Fifth International Workshop on Argumentation in Multi-Agent Systems*, pages 9–23. Springer-Verlag, 2008.
- [2] N. Belnap. Backwards and forwards in the modal logic of agency. *Philosophy and Phenomenological Research*, 51(777–807), 1991.
- [3] N. Belnap and M. Perloff. The way of the agent. *Studia Logica*, 51(3/4):463–484, 1992.
- [4] N. Belnap, M. Perloff, and M. Xu. *Facing the Future*. Oxford University Press, 2001.
- [5] K. E. Biggers and T. R. Ioerger. Automatic generation of communication and teamwork within multi-agent teams. *Applied Artificial Intelligence*, 15(10):875–916, 2001.
- [6] G. Boella and L. van der Torre.  $\Delta$ : The social delegation cycle. In *Proceedings of the 7th International Workshop on Deontic Logic in Computer Science*, volume 3065 of *Lecture Notes in Computer Science*, pages 29–42. Springer-Verlag, 2004.

- [7] G. Boella and L. van der Torre. Delegation of power in normative multiagent systems. In *Proceedings of the 8th International Workshop on Deontic Logic in Computer Science*, volume 4048 of *Lecture Notes in Computer Science*, pages 36–52. Springer-Verlag, 2006.
- [8] C. Castelfranchi. Commitments: From individual intentions to groups and organisations. In *Proceedings of the First International Conference on Multi-Agent Systems*, pages 41–48, 1995.
- [9] C. Castelfranchi and R. Falcone. Towards a theory of delegation for agent-based systems. *Robotics and Autonomous Systems*, 24:141–157, 1998.
- [10] B. F. Chellas. *Modal logic: An introduction*. Cambridge University Press, 1980.
- [11] B. F. Chellas. Time and modality in the logic of agency. *Studia Logica*, 51(3/4):485–517, 1992.
- [12] P. R. Cohen and H. J. Levesque. Communicative actions for artificial agents. In *Proceedings of the First International Conference on Multi-Agent Systems*, pages 65–72, 1995.
- [13] R. Conte and M. Paolucci. Responsibility for societies of agents. *Journal of Artificial Societies and Social Simulation*, 7(4), 2004. <http://jasss.soc.surrey.ac.uk/7/4/3.html>.
- [14] R. K. Dash, D. C. Parkes, and N. R. Jennings. Computational mechanism design: A call to arms. *IEEE Intelligent Systems*, 18(6):40–47, 2003.
- [15] E. Davis and L. Morgenstern. A first-order theory of communication and multi-agent plans. *Journal of Logic and Computation*, 15(5):701–749, 2005.
- [16] C. Dellarocas and M. Klein. Contractual agent societies: Negotiated shared context and social control in open multi-agent systems. In C. Dellarocas and R. Conte, editors, *Workshop on Norms and Institutions in MultiAgent Systems*, pages 1–11. ACM/AAAI, 2000.
- [17] F. Dignum, B. Dunin-Kępicz, and R. Verbrugge. Creating collective intention through dialogue. *Logic Journal of the IGPL*, 9(2):289–304, 2001.
- [18] M. Dummett. *Elements of intuitionism*. Clarendon Press, Oxford, UK, 1977.
- [19] D. Elgesem. The modal logic of agency. *Nordic Journal of Philosophical Logic*, 2(2):1–46, 1997.
- [20] R. Girle. Commands in dialogue logic. In D. M. Gabbay and H. J. Ohlbach, editors, *Practical Reasoning: Proceedings of the First International Conference on Formal and Applied Practical Reasoning*, volume 1455 of *Lecture Notes in Artificial Intelligence*, pages 246–260. Springer-Verlag, 1996.
- [21] G. Governatori and A. Rotolo. On the axiomatization of Elgesem’s logic of agency and ability. *Journal of Philosophical Logic*, 34(4):403–431, 2005.



- [22] D. Grossi, L. Royakkers, and F. Dignum. Organizational structure and responsibility: An analysis in a dynamic logic of organized collective agency. *Artificial Intelligence and Law*, 15:223–249, 2007.
- [23] C. L. Hamblin. *Imperatives*. Basil Blackwell, 1987.
- [24] H. L. A. Hart. Bentham on legal rights. In A. W. B. Simpson, editor, *Oxford Essays in Jurisprudence*, Second Series, pages 171–201. Oxford, 1973.
- [25] N. R. Jennings. Coordination techniques for distributed artificial intelligence. In G. M. P. O’Hare and N. R. Jennings, editors, *Foundations of Distributed Artificial Intelligence*, chapter 6, pages 187–210. John Wiley and Sons, 1996.
- [26] A. I. J. Jones and M. J. Sergot. A formal characterisation of institutionalised power. *Logic Journal of the IGPL*, 4(3):429–445, 1996.
- [27] A. Kuffik. Computers in control: Rational transfer of authority or irresponsible abdication of autonomy. *Ethics and Information Technology*, 1(3):173–184, 1999.
- [28] S. Kumar, M. J. Huber, D. R. McGee, P. R. Cohen, and H. J. Levesque. Semantics of agent communication languages for group interaction. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence*, pages 42–47, 2000.
- [29] N. Li, B. Grosz, and J. Feigenbaum. A practically implementable and tractable delegation logic. In *Proceedings of the Twenty-First IEEE Symposium on Security and Privacy*, pages 27–42. IEEE Computer Society Press, 2000.
- [30] L. Lindahl. *Position and change: A study in law and logic*. D. Reidel Publishing Company, Dordrecht, 1977.
- [31] E. Lorini, N. Troquard, A. Hertzog, and C. Castelfranchi. Delegation and mental states. In *Proceedings of the Sixth International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 622–624, 2007.
- [32] R. B. Marcus. Iterated deontic modalities. *Mind*, 75:580–582, 1966.
- [33] M. McCallum, W. W. Vasconcelos, and T. J. Norman. Organisational change through influence. *Autonomous Agents and Multi-Agent Systems*, 17(2):157–189, 2008.
- [34] L. Morgenstern. Knowledge preconditions for actions and plans. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, pages 867–874, 1987.
- [35] O. Morgenstern. Prolegomena to a theory of organization. ASTIA Document Number ATI 210734 (unpublished manuscript), 1951.
- [36] T. J. Norman and C. Reed. Group delegation and responsibility. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 491–498, 2002.

- [37] O. Pacheco and F. Santos. Delegation in a role-based organization. In A. Lomuscio and D. Nute, editors, *7th International Workshop on Deontic Logic in Computer Science*, volume 3065 of *Lecture Notes in Computer Science*, pages 209–227. Springer-Verlag, 2004.
- [38] J. S. Penberthy and D. Weld. UCPOP: A sound, complete, partial-order planner for ADL. In *Proceedings of the Third International Conference on Knowledge Representation and Reasoning*, pages 103–114, 1992.
- [39] I. Pörn. *The logic of power*. Basil Blackwell, 1970.
- [40] A. Prior. *Past, Present and Future*. Clarendon Press, Oxford, 1967.
- [41] S. D. Ramchurn, D. Huynh, and N. R. Jennings. Trust in multi-agent systems. *The Knowledge Engineering Review*, 19(1):1–25, 2004.
- [42] C. Reed and T. J. Norman. A formal characterisation of Hamblin’s action-state semantics. *Journal of Philosophical Logic*, 36:415–448, 2007.
- [43] N. Rescher. *The logic of commands*. Routledge & Kegan Paul Ltd, 1966.
- [44] M. P. Singh. Towards a formal theory of communication for multi-agent systems. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, pages 69–74, 1991.
- [45] M. P. Singh. A semantics for speech acts. *Annals of Mathematics and Artificial Intelligence*, 8:47–71, 1993.
- [46] M. P. Singh. *Multiagent Systems: A Theoretical Framework for Intentions, Know-how, and Communications*. Springer-Verlag, Secaucus, NJ, USA, 1994.
- [47] M. P. Singh, A. K. Chopra, and N. Desai. Commitment-based SOA. <http://people.engr.ncsu.edu/mpsingh/papers/drafts/Commitment-Based-SOA-TR-2007-19.pdf>, 2008.
- [48] I. A. Smith, P. R. Cohen, J. M. Bradshaw, M. Greaves, and H. Holmback. Designing conversation policies using joint intention theory. In *Proceedings of the Third International Conference on Multi-Agent Systems*, pages 269–276, 1998.
- [49] R. G. Smith and R. Davis. Frameworks for cooperation in distributed problem solving. *IEEE Transactions on Systems, Man and Cybernetics*, 11(1):61–70, 1981.
- [50] W. van der Hoek and M. Wooldridge. On the dynamics of delegation, cooperation and control: A logical account. In *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 701–708, 2005.
- [51] W. W. Vasconcelos, M. J. Kollingbaum, and T. J. Norman. Normative conflict resolution in multi-agent systems. *Journal of Autonomous Agents and Multi-Agent Systems*, 19(2):124–152, 2009.
- [52] G. H. von Wright. *An essay in deontic logic and the general theory of action*, volume 21 of *Acta philosophica Fennica*. North-Holland, Amsterdam, 1968.

- [53] D. N. Walton and E. C. W. Krabbe. *Commitment in dialogue: Basic concepts of interpersonal reasoning*. SUNY, New York, 1995.
- [54] D. N. Walton, C. Reed, and F. Macagno. *Argumentation Schemes*. Cambridge University Press, 2008.