# La confiance dans les relations entre agents : essai de modélisation en Logique Modale

R. Demolombe⋆
robert.demolombe@orange.fr

⋆Institut de Recherche en Informatique de Toulouse
Toulouse, France

**Résumé :**
La notion de confiance fait référence à plusieurs notions qui sont analysées de façon informelle, d'abord sur un exemple, puis dans le cas général. On propose ensuite un essai de formalisation en logique modale où l'on distingue les justifications et les motivations de la confiance. On montre que ces justifications peuvent être de nature soit empirique soit analytique, et que ces dernières font appel à la confiance dans d'autres propriétés : la capacité de créer un certain état des choses, la possibilité de déclencher des actions à partir d'intentions, le respect des normes, l'intérêt mutuel de deux agents, ou l'intérêt d'un agent pour l'autre.
La formalisation consiste à expliciter les relations entre les diverses modalités qui servent à définir ces notions. La relation entre antécédent et conséquent est exprimée par une logique des conditionnelles dont l'axiomatique est donnée en annexe.
Enfin, on montre comment ce cadre logique permet de comparer différentes définitions qui ont été exprimées en logique modale [1].

**Mots-clés :** Confiance, Logique Modale

**Abstract:**
The concept of trust refers to several concepts which are informally analyzed first in the context of an example, and then from an abstract point of view. An essay of formalization in modal logic is proposed where trust justifications and trust motivations are clearly distinguished. Justifications may be either empirical or analytical. It is shown that in the latter case they refer to trust into specific properties : the ability to reach some state of affairs, the possibility to trigger actions from intentions, norms fulfillment, mutual interest, or interest to fellow.
The formalization makes explicit the relationships between several modalities which are involved in the definitions of these properties. The entailment relationship is formalized in a minimal conditional logic plus some additional axiom schemas which are given in the annex.
Finally, it is shown how this logical frameworks allows to compare some distinct trust definitions that have been expressed in modal logic.

**Keywords:** Trust, Modal Logic

## 1 Introduction

Dans les relations entre agents la confiance joue un rôle très important, en particulier quand un agent compte sur d'autres agents pour réaliser un but, ou quand il doit faire des choix qui dépendent du comportement futur d'autres agents.

Les exemples abondent autant dans le domaine des relations individuelles que dans les relations sociales (l'économie est un exemple d'actualité assez convaincant), ou que dans les relations entre individus et agents institutionnels (par exemple entre une personne et sa banque).

Dans le domaine de l'informatique, ces relations peuvent utiliser comme intermédiaires des agents logiciels qui agissent pour le compte d'individus ou d'agents institutionnels. Par exemple, quand la gestion d'informations sensibles est confiée à des agents logiciels, se pose le problème de la confiance dans la préservation de la confidentialité, de l'intégrité et de la disponibilité d'accès à ces informations.

La multiplicité des travaux et des conférences consacrés à la confiance [2, 9, 6, 12, 16, 13] a montré qu'il s'agit d'une notion complexe dont la définition ne fait pas vraiment consensus. L'objectif de cet article est d'essayer de préciser les différentes notions qui interviennent dans la confiance. Pour cela, après une analyse informelle (section 2), on utilisera la Logique Modale pour leur donner un sens plus précis (section 3), sans chercher à détailler la formalisation jusqu'à un point qui rendrait la compréhension intuitive trop difficile. La dimension temporelle joue un rôle important dans la confiance et elle sera un peu plus analysée dans la section 4. L'analyse de la section 3 montre que la confiance peut se définir en termes de confiance dans des propriétés plus élémentaires qui sont récapitulées dans la section 5. Après avoir situé notre analyse par rapport aux travaux existants (section 6), on résumera en conclusion les principaux résultats et on proposera plusieurs directions pour des travaux futurs.

## 2 Analyse informelle

Avant d'aborder une analyse générale de la confiance nous allons étudier un exemple qui re-

1. Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-06-SETI-006.

présente bien les différentes questions qui seront étudiées ensuite d'un point de vue général.

Considérons un conducteur de voiture qui prévoit un voyage par autoroute, et qui envisage qu'une roue vienne à crever. Dans ce cas il devra se garer sur le bas côté et deux problèmes vont se poser : 1) réparer sa roue et 2) ne pas être accidenté par une voiture qui pourrait l'accrocher. Dans cette situation il aura deux buts : que la roue soit réparée (proposition $\phi$), et que la voiture reste dans l'état non accidentée (proposition $\phi'$).

Il se pose alors la question de savoir s'il peut avoir confiance dans le fait que, dans ces circonstances, sa roue sera réparée (proposition $\Diamond\phi$), et dans le fait que pendant tout ce temps il ne sera pas accidenté (proposition $\Box\phi'$). Supposons qu'il donne une réponse positive à ces deux questions.

Dans ce cas (F1), le conducteur (on l'appelle $i$) croit que si la roue est crevée (proposition $\neg\phi$) et que son but est qu'elle soit réparée (proposition $Goal_i\Diamond\phi$), alors la roue sera réparée.

D'autre part (M1), il croit que si la voiture n'est pas accidentée (proposition $\phi'$) et que son but est qu'elle reste non accidentée (proposition $Goal_i\Box\phi'$), alors elle restera non accidentée.

C'est le fait d'envisager la crevaison qui a **motivé** la question d'avoir, ou non, confiance. Une autre question est de savoir si cette confiance est **justifiée**.

Les justifications de ses croyances peuvent être de différentes natures :

- **Empiriques**. Dans les justifications empiriques on peut encore distinguer deux cas :
  - **Observations**. Par exemple, il est déjà arrivé plusieurs fois au conducteur de crever, et à chaque fois la roue a été réparée et il n'a pas été accidenté.
  - **Réputation**. Par exemple, le conducteur a entendu dire que tout le monde dit que si on crève on peut réparer sans être accidenté.
- **Analytiques**. Par exemple, le conducteur croit (F2) que dans ces circonstances il y aura quelqu'un d'autre (appelons-le $j$) qui est capable de réparer la roue et qui essaiera de réparer la roue. Et il croit aussi (M2) qu'aucun conducteur capable de l'accrocher n'essaiera de l'accrocher.

Ces justifications analytiques peuvent être considérées comme d'autres formes de confiance, qui elles mêmes peuvent avoir des justifications empiriques ou analytiques. Dans le cas des justifications analytiques on peut raffiner l'analyse de la façon suivante.

La confiance exprimée par la croyance (F2) peut être justifiée par le fait (F3) que $i$ croit que dans ces circonstances il y aura quelqu'un ayant l'intention de réparer sa roue qui transforme ses intentions en actions, et qui est capable de la réparer.

La confiance exprimée par la croyance (M2) peut être justifiée par le fait (M3) que $i$ croit que, dans ces circonstances, aucun conducteur capable de causer l'accident et qui transforme ses intentions en action n'aura l'intention de causer l'accident.

La confiance exprimée par la croyance (F3) peut être justifiée par le fait (F4.1) que $i$ croit que dans ces circonstances il y aura quelqu'un, $j$, qui croit qu'il doit réparer la roue de $i$, qui respecte les obligations (c'est-à-dire qui adopte l'intention de faire ce qu'il croit qu'il est obligatoire de faire), qui transforme ses intentions en actions, et qui est capable de réparer la roue. Dans ce cas la confiance de $i$ se fonde, entre autres, sur le fait que $j$ respecte les normes. On dira par la suite qu'un agent est "*obéissant*" s'il adopte l'intention de faire que l'on ait $\phi$ quand il croit qu'il est obligatoire qu'il adopte l'intention de faire que l'on ait $\phi$.

Bien que cette obligation ne soit pas dans le code de la route, il se peut que $i$ croit que $j$ croit que c'est une obligation morale.

Une autre éventualité est que la croyance (F3) est justifiée par le fait que (F4.2) $i$ croit que dans ces circonstances il y aura quelqu'un $j$ qui, en échange de la promesse d'une contrepartie, adoptera l'intention de réparer la roue. Cette personne $j$ peut-être, par exemple, un dépanneur auquel $i$ fait appel, et vis-à-vis duquel $i$ s'engage à payer l'intervention à condition que $j$ fasse la réparation, et $j$ s'engage à faire la réparation à condition que $i$ s'engage à payer l'intervention. Dans ce cas la confiance de $i$ est fondée sur l'intérêt mutuel de $i$ et de $j$.

Enfin, la confiance exprimée par la croyance (F3) peut être justifiée par le fait (F4.3) que $i$ croit que dans ces circonstances il y aura quelqu'un, $j$, qui croit que $i$ a pour but que sa roue soit réparée, et que cela suffit à $j$ pour adopter l'intention de réparer la roue de $i$. Il se peut que $j$ ait connaissance du but de $i$, soit parce que

$i$ fait appel à lui par un geste quelconque, soit parce que $j$ observe que $i$ est en train d'essayer de réparer la roue. Dans ce cas la confiance de $i$ se fonde sur le fait que l'attitude de $j$ est déterminée uniquement par l'intérêt qu'il porte à l'autre ($i$ en l'occurrence).

La confiance exprimée par la croyance (M3) peut être justifiée par le fait (M4.1) que $i$ croit que dans ces circonstances il n'y aura aucun agent $j$ qui croit qu'il est permis d'avoir l'intention de causer l'accident, qui ne fait que ce qui est permis (c'est-à'dire qui n'adopte pas l'intention de faire quelque chose quand il ne croit pas que cela est permis), qui transforme ses intentions en actions, et qui est capable de causer l'accident. On dira par la suite qu'un agent est "*honnête*" s'il n'adopte pas l'intention de faire que l'on ait $\phi$ quand il ne croit pas qu'il est permis de faire que l'on ait $\phi$.

La confiance exprimée par la croyance (M3) peut aussi être justifiée par le fait (M4.2) que $i$ croit que, dans ces circonstances, il n'y aura aucun agent $j$ qui croit qu'en demandant à un autre agent $k$ de causer l'accident moyennant une contre partie, cet agent $k$ s'engagera à causer l'accident à condition que $j$ s'engage à la contre partie, et $j$ s'engagera à la contre partie si $k$ s'engage à causer l'accident. Ici, la confiance de $i$ est fondée sur la croyance qu'aucun couple d'agents $j$ et $k$ n'a pour intérêt mutuel que l'accident ait lieu.

Enfin, la confiance exprimée par la croyance (M3) peut être justifiée par le fait (M4.3) que $i$ croit que dans ces circonstances il n'y aura aucun agent $j$ qui croit possible que $i$ ait pour but que l'accident ait lieu, et que cela suffit à $j$ pour adopter l'intention de causer l'accident, et qui transforme ses intention en action, et qui est capable de causer l'accident. Dans ce cas, comme pour (F4.3), la confiance de $i$ se fonde sur le fait que l'attitude de $j$ est déterminée uniquement par l'intérêt qu'il porte à l'autre.

Nous allons maintenant donner une présentation générale, mais informelle, des différentes sortes de confiance susceptibles d'être justifiées par d'autres formes de confiance. On adoptera dans cette présentation, un langage abrégé et semi-formel. On distinguera deux grands cas : celui où le but de $i$ est de la forme "faire que l'on ait $\phi$", et celui où il est de la forme "maintenir le fait qu'on ait $\phi$".

**Faire que l'on ait $\phi$**

(F1) L'agent $i$ croit que si $\neg\phi$ et $Goal_i\Diamond\phi$, alors dans le futur on aura $\phi$.

Dans le futur on aura $\phi$ s'il existe un agent $j$ tel que ($j$ est capable de faire en sorte que l'on ait $\phi$) ET ($j$ essaie de faire que l'on ait $\phi$).

(F2) L'agent $i$ croit que si $\neg\phi$ et $Goal_i\Diamond\phi$, alors il existe un agent $j$ tel que ($j$ est capable de faire en sorte que l'on ait $\phi$) ET ($j$ essaie de faire que l'on ait $\phi$).

$j$ essaie de faire que l'on ait $\phi$ si ($j$ essaie de faire que l'on ait $\phi$ quand il en a l'intention) ET ($j$ a l'intention de faire que l'on ait $\phi$).

(F3) L'agent $i$ croit que si $\neg\phi$ et $Goal_i\Diamond\phi$, alors il existe un agent $j$ tel que ($j$ essaie de faire que l'on ait $\phi$ quand il en a l'intention) ET ($j$ est capable de faire en sorte que l'on ait $\phi$) ET ($j$ a l'intention de faire que l'on ait $\phi$).

$j$ a l'intention de faire que l'on ait $\phi$ si :

- **Respect des normes** : $j$ croit qu'il est obligatoire que $j$ ait l'intention de faire que l'on ait $\phi$.
  Il est obligatoire que $j$ ait l'intention de faire que l'on ait $\phi$ si ($i$ demande à $j$ d'avoir l'intention de faire que l'on ait $\phi$) ET ($i$ a le pouvoir institutionnel de faire qu'il soit obligatoire que $j$ ait l'intention de faire que l'on ait $\phi$ en demandant à $j$ d'avoir l'intention de faire que l'on ait $\phi$).
- **Intérêt mutuel** : $j$ croit que ($i$ demande à $j$ de s'engager à faire que l'on ait $\phi$ à condition que $i$ s'engage à faire que l'on ait $\psi$) ET ($j$ s'engage à faire que l'on ait $\phi$ à condition que $i$ s'engage à faire que l'on ait $\psi$) ET ($i$ s'engage à faire que l'on ait $\psi$ à condition que $j$ s'engage à faire que l'on ait $\phi$).
- **Intérêt pour l'autre** : $j$ croit que $\neg\phi$ et $Goal_i\Diamond\phi$.
  $j$ croit que $Goal_i\Diamond\phi$ si :
  - $i$ demande à $j$ de faire que l'on ait $\phi$, ou
  - $j$ observe que ($\neg\phi$ ET $i$ essaie de faire que l'on ait $\phi$).

Pour illustrer ces trois possibilités on peut considérer le cas d'une personne $i$ dont le but est d'être transportée en voiture à une certaine destination. Si $i$ est un autostopeur, il a confiance dans le fait qu'en levant le bras il y aura un agent $j$, qui agit "pour l'intérêt de l'autre", qui acceptera de le prendre sans contrepartie. Par contre, si $i$ cherche un taxi, il a confiance dans le fait qu'en levant le bras un chauffeur de taxi $j$ s'arrêtera pour le transporter moyennant un enga-

gement mutuel. On notera que cela présuppose en plus que $i$ et $j$ aient confiance dans le fait que leurs engagements respectifs seront respectés. Ce qui ramène au cas "respect des normes". Dans un contexte similaire, si $i$ est un agent de police, il a le pouvoir institutionnel, en levant le bras, de créer l'obligation pour $j$ de s'arrêter, et il a confiance dans le fait que $j$ respectera l'obligation.

Les protocoles qui peuvent conduire à un engagement mutuel de $i$ et $j$ peuvent être très variés, celui qui a été présenté plus haut n'est donné qu'à titre d'exemple. Dans d'autres contextes il peut être fait appel à une tierce partie de confiance, comme un notaire. Le protocole peut même faire appel à deux autres agents qui représentent $i$ et $j$ et qui ont confiance l'un dans l'autre, comme dans le protocole de la lettre de crédit (voir [1]) utilisé pour le commerce maritime.

**Maintenir le fait qu'on ait** $\phi$

(M1) L'agent $i$ croit que si $\phi$ et $Goal_i\Box\phi$, alors dans le futur on aura toujours $\phi$.

L'agent $i$ croit que dans le futur on aura toujours $\phi$ si aucun agent $j$ capable de faire que l'on ait $\Diamond\neg\phi$ n'essaie de faire que l'on ait $\Diamond\neg\phi$.

(M2) L'agent $i$ croit que si $\phi$ et $Goal_i\Box\phi$, alors il n'existe aucun agent $j$ tel que ($j$ est capable de faire que l'on ait $\neg\phi$) ET ($j$ essaie de faire que l'on ait $\neg\phi$).

L'agent $i$ croit que dans le futur on aura toujours $\phi$ si aucun agent $j$ tel que ($j$ essaie de faire que l'on ait $\neg\phi$ quand il en a l'intention) et ($j$ est capable de faire que l'on ait $\neg\phi$) n'a l'intention de faire que $\Diamond\neg\phi$.

(M3) L'agent $i$ croit que si $\phi$ et $Goal_i\Box\phi$, alors il n'existe aucun agent $j$ tel que ($j$ essaie de faire que l'on ait $\neg\phi$ quand il en a l'intention) ET ($j$ est capable de que l'on ait $\neg\phi$) ET ($j$ a l'intention de que l'on ait $\neg\phi$).

$j$ n'a pas l'intention de faire $\neg\phi$ si :

- **Respect des normes** : $j$ ne croit pas qu'il soit permis que $j$ ait l'intention de que l'on ait $\neg\phi$.
- **Intérêt mutuel** : il est faux que $j$ croit qu'il existe $k$ tel que : ($k$ demande à $j$ de s'engager à faire que $\neg\phi$ à condition que $k$ s'engage à que l'on ait $\psi$) ET ($j$ s'engage à que l'on ait $\neg\phi$ à condition que $k$ s'engage à que l'on ait $\psi$) ET ($k$ s'engage à que l'on ait $\psi$ à condition que $j$ s'engage à que l'on ait $\neg\phi$)
- **Intérêt pour l'autre** : $j$ ne croit pas que $\phi$ et $Goal_i\Diamond\neg\phi$.
  $j$ ne croit pas que $Goal_i\Diamond\neg\phi$ si :
  - $i$ ne demande pas à $j$ de que l'on ait $\neg\phi$, ou
  - $j$ n'observe pas que ($\phi$ ET $i$ essaie de que l'on ait $\neg\phi$).

On notera que, dans le cas "intérêt mutuel", $j$ ne croit pas que quelqu'un ($k$) le "payera" pour faire échouer le but de $i$. Bien que formellement $k$ puisse être $i$ lui-même, il n'est pas vraisemblable que $j$ croie que $i$ puisse "payer" $j$ pour qu'il mette en échec son propre but.

## 3 Essai de formalisation

Nous allons introduire certains opérateurs modaux pour décrire les différentes attitudes mentales que nous venons de voir. La signification de ces opérateurs sera donnée de façon purement intuitive, car ce qui nous intéresse en premier est de montrer que les différentes formes de confiance s'expriment par des relations entre les notions représentées par ces opérateurs.

Le seul opérateur dont l'axiomatique sera partiellement explicitée est l'opérateur qui exprime le fait que "si on a $\phi$, alors on a $\psi$", ou "$\phi$ entraîne $\psi$". On sait que si on utilise l'implication matérielle pour représenter la relation entre $\phi$ et $\psi$ on obtient des conséquences paradoxales qui tiennent au fait que l'implication matérielle est vraie quand l'antécédent est faux. Quand elle est dans le champ d'un opérateur modal on peut éviter ce type de paradoxe, mais c'est au prix de conditions compliquées, qui n'ont rien de naturel, et servent plutôt d'"astuces" pour éviter les paradoxes qu'à exprimer des conditions intuitives. C'est pour ces raisons que nous avons choisi d'exprimer la relation entre antécédent et conséquent avec un opérateur de "conditionnelle" (voir "minimal conditional logic" dans [3] section 10.1) auquel on n'impose que les propriétés dont on a besoin ici. On peut lire intuitivement ces opérateurs de la façon suivante.

$\phi \Rightarrow \psi$ : $\phi$ entraîne $\psi$.
$Bel_i\phi$ : $i$ croit que $\phi$.
$Goal_i\phi$ : $i$ a pour but que l'on ait $\phi$.
$\Box\phi$ : $\phi$ est vraie maintenant et dans tous les instants futurs.
$\Diamond\phi$ : il y a un instant, maintenant ou dans le futur, où $\phi$ est vraie.
$\Diamond\phi \stackrel{\text{def}}{=} \neg\Box\neg\phi$.
$Attempt_i\phi$ : $i$ essaie de faire en sorte que l'on ait $\phi$.

$Int_i\phi$ : $i$ a l'intention de faire que l'on ait $\phi$.
$Obg\phi$ : il est obligatoire que l'on ait $\phi$.
$Perm\phi$ : il est permis que l'on ait $\phi$.
$Perm\phi \stackrel{\text{def}}{=} \neg Obg\neg\phi$.
$Ask_{i,j}\phi$ : $i$ demande à $j$ de faire que l'on ait $\phi$.
$Commit_i(\phi \mid \psi)$ : $i$ s'est engagé à faire que l'on ait $\phi$ à condition que l'on ait $\psi$.
$Obs_i\phi$ : $i$ observe que l'on a $\phi$.

Nous allons maintenant exprimer les différentes attitudes des agents qui ont été présentées dans la section précédente en utilisant ces modalités. Pour ne pas avoir des formules trop complexes nous n'exprimerons pas tous les détails liés à la dimension temporelle. On verra dans la section suivante comment on peut aller plus loin dans ce sens.

**Faire que l'on ait $\phi$**

(F1) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow \Diamond\phi)$

On définit la notion de capacité à faire que l'on ait $\phi$ de la façon suivante.
$Able_j\phi \stackrel{\text{def}}{=} Attempt_j\phi \Rightarrow \Diamond\phi$

(F2) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow \exists j(Able_j\phi \wedge Attempt_j\phi))$ [2]

Le fait qu'un agent a la propriété de traduire son intention en action est exprimé par la définition suivante.
$PosActive_j\phi \stackrel{\text{def}}{=} Int_j\phi \Rightarrow Attempt_j\phi$

On notera qu'on a : $(Able_j\phi \wedge Attempt_j\phi) \Rightarrow \Diamond\phi$, d'où :

(F3) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow \exists j(PosActive_j\phi \wedge Able_j\phi \wedge Int_j\phi))$

Le fait qu'un agent est obéissant est exprimé par la propriété suivante.
$Obey_j\phi \stackrel{\text{def}}{=} Bel_jObgInt_j\phi \Rightarrow Int_j\phi$
On notera que même si on a : $(Bel_jObgInt_j\phi) \wedge Obey_j\phi$, cela garantit seulement que $j$ a l'intention de faire que l'on ait $\phi$. On a alors :

(F4.1) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow \exists j(Obey_j\phi \wedge PosActive_j\phi \wedge Able_j\phi \wedge Bel_jObgInt_j\phi))$

Le fait que, du point de vue de $j$, $i$ a le pouvoir, dans l'institution $s$, de créer l'obligation pour $j$ d'avoir l'intention de faire que l'on ait $\phi$ est défini de la façon suivante. [3]
$InstPower_{i,j}\phi \stackrel{\text{def}}{=} Ask_{i,j}\phi \Rightarrow_s Bel_jObgInt_j\phi$
On a alors :

(F5.1) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow \exists j(InstPower_{i,j}\phi \wedge Obey_j\phi \wedge PosActive_j\phi \wedge Able_j\phi \wedge Ask_{i,j}\phi))$

Dans l'axiomatique des conditionnelles mentionnée en annexe on peut montrer facilement qu'on a : (F2) implique (F1), (F3) implique (F1), (F4.1) implique (F1), et (F5.1) implique (F1).

On définit le fait que $i$ et $j$ se sont mutuellement engagés, à la demande de $i$, à faire que l'on ait $\phi$ à condition que l'autre fasse que l'on ait $\psi$ de la façon suivante.
$Contract_{i,j}(\phi, \psi) \stackrel{\text{def}}{=}$
$InstPower_{i,j}(Commit_j(\phi \mid Commit_i\psi)) \wedge$
$Obey_j(Commit_j(\phi \mid Commit_i\psi)) \wedge$
$(Commit_j(\phi \mid Commit_i\psi) \Rightarrow$
$Commit_i(\psi \mid Commit_j\phi)) \wedge$
$(Commit_i(\psi \mid Commit_j\phi) \Rightarrow Int_j\phi)$
On a alors :

(F4.2) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow \exists j(Contract_{i,j}(\phi, \psi) \wedge PosActive_j\phi \wedge Able_j\phi \wedge Ask_{i,j}Commit_j(\phi \mid Commit_i\psi))$

On a : (F4.2) implique (F1).

On définit le fait que $j$ est disposé à satisfaire le but de $i$ sans contrepartie de la façon suivante :
$ActAltruis_{j,i}\phi \stackrel{\text{def}}{=}$
$Bel_j(\neg\phi \wedge Goal_i\Diamond\phi) \Rightarrow Int_j\phi$
On a alors :

(F4.3) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow \exists j(Act Altruis_{j,i}\phi \wedge PosActive_j\phi \wedge Able_j\phi \wedge Bel_j(\neg\phi \wedge Goal_i\Diamond\phi)))$

On a (F4.3) implique (F1).

**Maintenir le fait qu'on ait $\phi$**

(M1) $Bel_i(\phi \wedge Goal_i\Box\phi \Rightarrow \Box\phi)$

(M2) $Bel_i(\phi \wedge Goal_i\Box\phi \Rightarrow \neg\exists j(Able_j\neg\phi \wedge Attemt_j\neg\phi))$

2. Pour éviter un trop grand nombre de parenthèses on utilise la formule : $\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow \exists j(Able_j\phi \wedge Attempt_j\phi)$, à la place de : $(\neg\phi \wedge Goal_i\Diamond\phi) \Rightarrow (\exists j(Able_j\phi \wedge Attempt_j\phi))$.

3. On peut juger cette définition exagérément simplifiée. On devrait avoir plutôt : $InstPower_{i,j,s}\phi \stackrel{\text{def}}{=} Bel_j(Ask_{i,j,s}\phi \Rightarrow_s Obg_sInt_j\phi)$, c'est-à-dire : $j$ croit que $i$ a le pouvoir, dans l'institution $s$, de créer l'obligation $Obg_sInt_j\phi$ en demandant à $j$ de faire que l'on ait $\phi$. Voir [8] pour les actes de communication qui ont des effets institutionnels.

Si un agent capable de que l'on ait $\Diamond\neg\phi$ essaie de que l'on ait $\Diamond\neg\phi$, alors on a $\Diamond\neg\phi$.

Formellement, on a :
$(Able_j\neg\phi \wedge Attemt_j\neg\phi) \Rightarrow \Diamond\neg\phi$.

Un agent $i$, peut accepter, en plus, que si on a $\phi$ et aucun agent capable de faire que l'on ait $\neg\phi$ n'essaie de faire que l'on ait $\neg\phi$, alors on a toujours $\phi$. Ceci revient à accepter que $\neg\phi$ ne peut résulter que de l'action d'un agent capable de faire que l'on ait $\neg\phi$. Cette croyance de $i$ est exprimée par :
(PERSIST1) $Bel_i(\phi \wedge \neg\exists j(Able_j\neg\phi \wedge Attemt_j\neg\phi) \Rightarrow \Box\phi)$

On a (M2) et (PERSIST1) implique (M1).

Le fait qu'un agent est capable de se retenir de faire $\phi$, c'est-à-dire de ne pas essayer de faire que l'on ait $\phi$ s'il n'en a pas l'intention est défini par :
$RefActive_j\phi \stackrel{\text{def}}{=} \neg Int_j\phi \Rightarrow \neg Attempt_j\phi$
Formellement on a :
$\neg Int_j\neg\phi \wedge RefActive_j\neg\phi \rightarrow \neg Attempt_j\neg\phi$.
On a alors :

(M3) $Bel_i(\phi \wedge Goal_i\Box\phi \Rightarrow \neg\exists j(RefActive_j\neg\phi \wedge Able_j\neg\phi \wedge Int_j\neg\phi))$

Un agent $i$, peut accepter, en plus, que si aucun agent qui peut se retenir d'essayer de que l'on ait $\neg\phi$ et qui est capable de que l'on ait $\neg\phi$ n'a l'intention de que l'on ait $\neg\phi$, alors on a toujours $\phi$. Soit :
(PERSIST2) $Bel_i(\phi \wedge \neg\exists j(RefActive_j\neg\phi \wedge Able_j\neg\phi \wedge Int_j\neg\phi) \Rightarrow \Box\phi)$

Cette hypothèse est risquée car il pourrait y avoir des agents capables de faire $\neg\phi$, et qui essaient de faire que $\neg\phi$ bien qu'ils n'en aient pas l'intention. Dans l'exemple de la section précédente, ce pourrait être le cas d'un conducteur $j$ qui accroche la voiture de $i$ involontairement.

On peut noter que (PERSIST2) est équivalent à :

$Bel_i(\phi \wedge \forall j(RefActive_j\neg\phi \wedge Able_j\neg\phi \rightarrow \neg Int_j\neg\phi) \Rightarrow \Box\phi)$

On a (M3) et (PERSIST2) implique (M1).

Le fait que $j$ est honnête est défini par :
$Honest_j\phi \stackrel{\text{def}}{=} \neg Bel_jPermInt_j\phi \Rightarrow \neg Int_j\phi$
On a alors :

(M4.1) $Bel_i(\phi \wedge Goal_i\Box\phi \Rightarrow \neg\exists j(Honest_j\neg\phi \wedge PosActive_j\neg\phi \wedge Able_j\neg\phi \wedge \neg Bel_jPermInt_j\neg\phi))$

Formellement on a :
$\neg Bel_jPermInt_j\neg\phi \wedge Honest_j\neg\phi \rightarrow \neg Int_j\neg\phi$.

Un agent $i$, peut accepter, en plus, que si aucun agent qui est honnête pour $\neg\phi$, qui peut se retenir d'essayer de que l'on ait $\neg\phi$ et qui est capable de que l'on ait $\neg\phi$, ne croit qu'il est permis d'avoir l'intention de faire que $\neg\phi$, alors on a toujours $\phi$. Soit :
(PERSIST3.1) $Bel_i(\phi \wedge \neg\exists j(Honest_j\neg\phi \wedge PosActive_j\neg\phi \wedge Able_j\neg\phi \wedge \neg Bel_jPermInt_j\neg\phi) \Rightarrow \Box\phi)$

Cette hypothèse est risquée car il peut y avoir des agents qui peuvent se retenir d'essayer de que l'on ait $\neg\phi$ et qui sont capables de que l'on ait $\neg\phi$, mais qui ne sont pas honnêtes.

On a (M4.1) et (PERSIST3.1) implique (M1).

(M4.2) $Bel_i(\phi \wedge Goal_i\Box\phi \Rightarrow \neg\exists j(Bel_j(\exists k((Contract_{k,j}(\neg\phi,\psi) \wedge PosActive_j\neg\phi \wedge Able_j\neg\phi \wedge Ask_{k,j}Commit_j(\neg\phi \mid Commit_j\psi))))))$

Un agent $i$ peut accepter en plus une hypothèse similaire à (PERSIST3.1). Soit :
(PERSIST3.2) $Bel_i(\phi \wedge \neg\exists j(Bel_j(\exists k((Contract_{k,j}(\neg\phi,\psi) \wedge PosActive_j\neg\phi \wedge Able_j\neg\phi \wedge Ask_{k,j}Commit_j(\neg\phi \mid Commit_j\psi))))) \Rightarrow \Box\phi)$

On a (M4.2) et (PERSIST3.2) implique (M1).

Le fait qu'un agent $j$ se retient d'adopter l'intention de faire que l'on ait $\phi$ s'il croit qu'il est possible que le but de $i$ est qu'il n'adopte pas l'intention de faire que l'on ait $\phi$ est défini par :
$RefAltruis_{i,j}\phi \stackrel{\text{def}}{=} \neg Bel_j\neg Goal_i\neg Int_j\phi \Rightarrow \neg Int_j\phi$
On a alors :

(M4.3) $Bel_i(\phi \wedge Goal_i\Box\phi \Rightarrow \neg\exists j(RefAltruis_{i,j}\neg\phi \wedge PosActive_j\neg\phi \wedge Able_j\neg\phi \wedge \neg Bel_j\neg Goal_i\neg Int_j\neg\phi))$

Un agent $i$ peut en plus adopter une hypothèse similaire à à (PERSIST3.1). Soit :
(PERSIST3.3) $Bel_i(\phi \wedge \neg\exists j(RefAltruis_{i,j}\neg\phi \wedge PosActive_j\neg\phi \wedge Able_j\neg\phi \wedge \neg Bel_j\neg Goal_i\neg Int_j\neg\phi) \Rightarrow \Box\phi)$

On a (M4.3) et (PERSIST3.3) implique (M1).

La formalisation qui a été présentée jusqu'ici pourrait être raffinée dans plusieurs directions. D'abord en ce qui concerne les quantificateurs qui apparaissent dans les formules telles que (F2) et (M2). Si on s'inspire des logiques modales du premier ordre, la position des quantificateurs par rapport à la modalité $Bel_i$ permet de distinguer le fait que $i$ connaît, ou ne connaît pas, les agents $j$ qui son capables de faire que l'on ait $\phi$ ou de que l'on ait $\neg\phi$.

Par exemple, la formule (F'2) :
(F2) $\exists j Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow (Able_j\phi \wedge Attempt_j\phi))$
exprime que $i$ connaît un $j$ qui satisfait la propriété dans le champs de $Bel_i$, alors que ce n'est pas nécessairement le cas dans (F2).

Le cas de (M2) est plus délicat car il contient implicitement un quantificateur universel. En effet (M2) est équivalent à (M'2) :
(M'2) $Bel_i(\phi \wedge Goal_i\Box\phi \Rightarrow \forall j(Able_j\neg\phi \rightarrow \neg Attemt_j\neg\phi))$
La même transformation que celle que nous avons faite pour passer de (F2) à (M2) conduirait à (M"2) :
(M"2) $\forall j Bel_i(\phi \wedge Goal_i\Box\phi \Rightarrow (Able_j\neg\phi \rightarrow Attemt_j\neg\phi))$
Mais (M"2) est très problématique car elle suppose implicitement que $i$ connaît l'ensemble $I$ de tous les agents capables de faire que l'on ait $\neg\phi$ [4], et il y a de nombreux domaines d'application où ce n'est pas le cas. Par exemple, si $i$ a confié à certains agents logiciels la gestion d'informations confidentielles, il sait que ce sont ces agents qui ont la capacité de communiquer ces informations, et il peut avoir confiance, ou pas, dans le fait qu'aucun d'eux ne les communiquera à des personnes non autorisées. Par contre, si $i$ transmet ses informations via un réseau tel qu'Internet, il n'est pas très réaliste d'envisager que $i$ connaisse tous les agents qui ont la capacité d'altérer l'intégrité de ces informations.

Les différentes modalités que nous avons utilisées ont fait l'objet de nombreux travaux de formalisation, et si certaines, comme la modalité épistémique $Bel_i$ font l'objet de consensus, ce n'est pas le cas pour les modalités de but et d'intention $Goal_i$ et $Int_j$, et encore moins pour les modalités déontiques $Obg$ et $Perm$, ainsi que pour la modalités d'action $Attempt_j$.

Dans notre esprit il est essentiel dans le cadre de la confiance que la modalité $Attempt_j$ exprime la causalité, c'est-à-dire que si après que $j$ ait essayé de faire que l'on ait $\phi$ on a $\phi$, alors on a $\phi$ à cause de ce qu'a fait $j$. Même s'il est bien connu que la formalisation de la causalité pose de gros problèmes, en particulier à cause de la condition contre factuelle, les travaux de G. H. von Wright [18], de I. Pörn [17] ou de N. Belnap et J. Horty [10], peuvent servir de référence.

Malheureusement l'expression "essayer de faire que" laisse entendre que $j$ agit intentionnellement. Ce n'est pas le sens que nous voulons donner à $Attempt_j\phi$, et **on ne suppose pas** que : $Attempt_j\phi \Rightarrow Int_j\phi$. Par exemple, si le conducteur $i$ a confiance dans le fait qu'aucun autre conducteur $j$ ne va l'accrocher, l'ensemble $I$ mentionné plus haut contient non seulement les agents qui pourraient l'accrocher intentionnellement, mais aussi ceux qui pourraient le faire involontairement.

On notera que parmi les propriétés élémentaires qui peuvent faire l'objet de confiance ne figure pas le pouvoir institutionnel. En effet, la conditionnelle qui définit ce pouvoir exprime une norme, définie par une institution, qui relie l'antécédent et le conséquent [11]. Cette relation n'exprime pas la régularité de l'attitude d'un agent, et elle est de nature différente.

Enfin, faut-il envisager un cadre logique unique qui intégrerait une formalisation détaillée de toutes les modalités que nous utilisons. Mis à part le fait qu'un travail d'une telle ampleur constitue un défi considérable, le résultat serait d'une telle complexité que le bénéfice en terme de compréhension intuitive est loin d'être acquis. De plus, on pourrait toujours trouver des points insuffisamment formalisés, et comme le montre très bien J. Ladrière dans [14] il faut accepter des limites à la formalisation. Ici, nous nous sommes limité au niveau qui permet de comparer différents points de vue sur la formalisation de la confiance.

# 4 Dimension temporelle

Jusqu'à présent nous n'avons fait intervenir le temps que pour distinguer le cas où le but d'un agent $i$ est qu'advienne une situation où on a $\phi$, représenté par $\neg\phi \wedge \Diamond\phi$, et le cas où son but est

4. La formule (M"2) soulève aussi la question de savoir si elle est "domain independent", au sens défini dans [4], car, à notre connaissance, il n'y a aucune étude portant sur les formules "domain independent" en logique modale, et il faudrait s'assurer que la sous-formule $Able_j\neg\phi$ peut être considérée comme un prédicat qui définit le champ du quantificateur universel $\forall j$. La même question se pose pour le quantificateur existentiel $\exists j$ dans (F2).

que persiste une situation où on a $\phi$, représenté par $\phi \wedge \Box\phi$.

Il y a une autre distinction qui vaut la peine d'être analysée, car elle a conduit à des définitions différentes de la confiance, et peut-être à des confusions, c'est la distinction entre, d'une part, le cas où survient un événement, par exemple une roue qui crève, qui motive l'adoption à cet instant-là d'un but, et, éventuellement, l'adoption de la confiance dans le fait que le but sera atteint, et, d'autre part, le cas où, comme on l'a décrit dans la section 2, un agent envisage qu'à n'importe quel instant futur pourrait survenir le même événement. Dans ce cas c'est à l'instant présent qu'il doit décider, ou non, d'adopter la confiance dans le fait que son but sera satisfait quel que soit l'instant futur où l'événement aura lieu.

Nous allons voir comment cette distinction s'exprime en logique modale.

On utilisera les notations suivantes.

$Until(\phi)\psi$ : $\psi$ est vraie jusqu'à ce que $\phi$ soit vraie.
$Before(\phi)\psi$ : $\psi$ sera vraie à un certain instant avant que $\phi$ soit vraie.

**But motivé par un événement présent**

**Faire que l'on ait $\phi$**

En pratique $i$ veut que son but soit satisfait dans un certain délai, c'est-à-dire avant qu'une proposition représentée par $\delta$ soit vraie. Par exemple, $i$ veut que la roue soit réparée avant qu'il fasse nuit. D'autre part, le but de $i$ ne peut pas être satisfait à l'instant même où il a été adopté. Donc le but de $i$ persiste jusqu'à ce qu'il soit satisfait ou que le délai soit dépassé.

Si de plus $i$ adopte, à l'instant présent, la croyance dans le fait que dans toute situation telle que la situation présente le but sera satisfait avant $\delta$, on peut dire qu'il a confiance dans le fait que son but sera réalisé. Formellement la situation est représentée par :

$Bel_i(\neg\phi \wedge Until(\phi \vee \delta)Goal_i Before(\delta)\phi \Rightarrow Before(\delta)\phi)$

On notera que dans ce cas la relation entre le but et le fait qu'il sera satisfait s'exprime par une conditionnelle bien que la croyance de $i$ ne soit motivée que par l'événement présent. En effet, si $i$ adopte cette croyance c'est parce qu'il croit qu'il y a une relation de dépendance entre son but et le fait qu'il soit satisfait, et il n'est pas nécessaire qu'il croit que cette relation persiste au delà de l'instant où son but actuel est satisfait.

Ce serait une erreur de représenter la relation entre le but et sa satisfaction par une conjonction, bien qu'à partir de la confiance $i$ puisse **déduire** qu'il a le but **et** qu'il sera satisfait [5] .

**Maintenir le fait qu'on ait $\phi$**

Ici aussi, en pratique, le but de $i$ (ne pas être accidenté) n'est pas illimité dans le temps. Par exemple, son but peut être ne pas être accidenté jusqu'à ce qu'il puisse repartir, c'est-à-dire jusqu'à ce que la roue soit réparée. D'autre part, son but de ne pas être accidenté persiste jusqu'à ce qu'il puisse repartir. Si on représente par $\delta$ la limite qu'il accepte pour son but, on a alors dans le cas général :

$Bel_i(\phi \wedge Until(\delta)Goal_i Until(\delta)\phi \Rightarrow Until(\delta)\phi)$

**But motivé par un éventuel événement futur**

**Faire que l'on ait $\phi$**

Dans ce cas $i$ peut adopter le but à n'importe quel instant dans le futur, et il croit que s'il adopte le but, alors le but sera atteint. Cependant, ici aussi, en pratique $i$ n'envisage pas un futur illimité et sa confiance ne peut pas être illimitée dans le temps. Par exemple, il a confiance dans le fait que sa roue sera réparée si elle crève à n'importe quel instant de son voyage jusqu'à ce qu'il soit arrivé à son terme.

Dans le cas général, la confiance de $i$ s'exprime alors de la façon suivante.

$Bel_i(Until(\delta')(\neg\phi \wedge Goal_i Before(\delta)\phi \Rightarrow Before(\delta)\phi))$

**Maintenir le fait qu'on ait $\phi$**

Pour des raisons similaires, ici aussi le but et la confiance de $i$ ont des limites temporelles. Par exemple, $i$ croit que jusqu'à ce qu'il soit arrivé au terme de son voyage ($\delta'$), s'il est garé le long d'un autoroute et que son but est de ne pas se faire accrocher jusqu'à ce qu'il ait réparé ($\delta$), alors il ne sera pas accroché jusqu'à ce qu'il ait réparé.

5. On notera que si la relation entre l'antécédent et le conséquent était représentée par l'implication matérielle, c'est-à-dire par une formule de la forme : $Bel_i(A \rightarrow C)$, on aurait $Bel_i(A \rightarrow C) \wedge Bel_i A$ logiquement équivalent à $Bel_i(A \wedge C)$, alors que ce n'est pas le cas pour $Bel_i(A \Rightarrow C) \wedge Bel_i A$

Dans le cas général, la confiance de $i$ s'exprime alors de la façon suivante.

$Bel_i(Until(\delta')(\phi \wedge Goal_i Until(\delta)\phi \Rightarrow Until(\delta)\phi))$

## 5 Confiance dans les propriétés élémentaires

De l'analyse de la section 3 on peut conclure que la confiance dans le fait qu'un but soit atteint peut être justifiée par la confiance dans un certain nombre de propriétés élémentaires qui sont récapitulées ci-dessous. Ces propriétés s'expriment toutes sous forme de conditionnelles.

Du point de vue des relations temporelles entre antécédents et conséquents nous avons fait certaines hypothèses. On suppose que les relations entre attitudes mentales sont simultanées (relations entre le fait que $j$ croit, ou ne croit pas, quelque chose et le fait que $j$ adopte, ou non, l'intention de faire que l'on ait $\phi$), et on suppose que le passage d'une intention à une action, ou d'une action à l'effet de l'action, nécessite un délai dans le temps. On a alors :

$Able_j\phi \stackrel{\text{def}}{=} Attempt_j\phi \Rightarrow \Diamond\phi$
$PosActive_j\phi \stackrel{\text{def}}{=} Int_j\phi \Rightarrow \Diamond Attempt_j\phi$
$RefActive_j\phi \stackrel{\text{def}}{=} \neg Int_j\phi \Rightarrow \neg\Diamond Attempt_j\phi$
$Obey_j\phi \stackrel{\text{def}}{=} Bel_j ObgInt_j\phi \Rightarrow Int_j\phi$
$ActAltruis_{j,i}\phi \stackrel{\text{def}}{=} Bel_j(\neg\phi \wedge Goal_i\Diamond\phi) \Rightarrow Int_j\phi$
$Honest_j\phi \stackrel{\text{def}}{=} \neg Bel_j PermInt_j\phi \Rightarrow \neg Int_j\phi$
$RefAltruis_{i,j}\phi \stackrel{\text{def}}{=} \neg Bel_j(\phi \wedge Goal_i\Box\phi) \Rightarrow \neg Int_j\phi$

En général, la confiance est une propriété que $i$ attribue à $j$ pour les instants futurs. Mais, comme on l'a vu, en pratique cette confiance n'est pas illimitée dans le temps. De plus, en général, elle est restreinte à certains contextes, et ceci de deux façons différentes. D'une part, ce n'est que dans certains contextes que $i$ accordera sa confiance à $j$ pour une certaine durée dans le temps, et d'autre part, à un instant futur donné, la propriété attribuée à $j$ peut être restreinte aux cas où certaines conditions sont satisfaites à cet instant-là. Si on appelle $Prop_{i,j}\phi$ la propriété attribuée par $i$ à $j$, on a la forme générale de la confiance représentée par [6] :

$TrustProp_{i,j}\phi \stackrel{\text{def}}{=} Bel_i(BelContext \Rightarrow Until(\delta)(PropContext \Rightarrow Prop_{i,j}\phi))$

Par exemple, dans le cas de la confiance dans la capacité ou dans l'honnêteté de $j$ on a :

$TrustAble_{i,j}\phi \stackrel{\text{def}}{=} Bel_i(BelAContext \Rightarrow Until(\delta)(AContext \Rightarrow (Attempt_j\phi \Rightarrow \Diamond\phi)))$

$TrustHonest_{i,j}\phi \stackrel{\text{def}}{=} Bel_i(BelHContext \Rightarrow Until(\delta)(HContext \Rightarrow (\neg Bel_j PermInt_j\phi \Rightarrow \neg Int_j\phi)))$

## 6 Comparaison avec d'autres travaux

Nous nous restreindrons ici aux travaux qui proposent des formalisations en logique modale. D'une manière générale ils s'accordent sur le fait que la confiance est une attitude mentale qui s'exprime par une croyance d'un agent $i$ au sujet de propriétés attribuées à un autre agent $j$ [2, 9, 12, 6].

Ces propriétés peuvent-elles être étendues à des objets, ou à des systèmes physiques ? Par exemple, il se peut que, dans le cas où $i$ est un conducteur, $i$ déclare qu'il a confiance dans le fait que son régulateur automatique de vitesse se débranchera s'il appuie sur le bouton approprié. Nous pensons qu'en fait cette confiance ne concerne pas les systèmes physiques eux-mêmes. En effet, leurs comportements sont déterminés, d'une part, par les lois de la nature, et celles-ci ne peuvent faire l'objet d'un jugement subjectif tel que la confiance, et d'autre part, par les personnes qui les ont conçus, fabriqués ou utilisés. Ce sont donc, plus vraisemblablement, ces personnes qui peuvent faire, ou pas, l'objet de la confiance de $i$. On est donc ramené au cas précédent.

Les divergences les plus importantes portent sur les notions qui doivent être considérées comme constitutives de la confiance. En particulier, pour C. Castelfranchi et al. [2, 9] le but de $i$ est un élément constitutif de la confiance. Selon cette approche, la croyance de $i$ dans certaines propriétés de $j$ ne peut être considérée comme de la confiance que si cette croyance est motivée par le fait que $i$ a un certain but, et que celui-ci

6. Une notation, plus lourde, mais plus rigoureuse devrait être de la forme : $TrustProp_{i,j}(\phi, BelContext, PropContext, \delta)$.

pourrait être satisfait par $j$ en mettant en oeuvre cette propriété. Cette approche peut être formalisée en logique en exprimant que si $i$ croit que $j$ a cette propriété, alors nécessairement $i$ a un but qui peut être atteint grace à $j$. Cette contrainte peut s'exprimer, sans détailler tous les aspects temporels, par :

$TrustProp_{i,j}\phi \rightarrow Bel_i \Diamond Goal_i \phi$

On notera qu'ici la relation entre l'antécédent est le conséquent est exprimée par l'implication matérielle, et non par une conditionnelle, car la contrainte exprime une disjonction : ou bien on n'a pas $TrustProp_{i,j}\phi$, ou bien on a $Bel_i \Diamond Goal_i \phi$.

Pour d'autre auteurs, en particulier pour A.J.I. Jones [12], le but n'est pas considéré comme une condition nécessaire pour qu'on puisse appeler "confiance" certaines croyances, même s'il accepte que très souvent c'est le cas. Il considère que c'est essentiellement la régularité de l'attitude de $j$ qui constitue la notion de confiance. Il distingue deux types de régularité : la capacité à créer une certaine situation, et le respect de certaines obligations. Il exprime la première par la formule : $Bel_i(\psi \rightarrow E_j\phi)$, et la seconde par : $Bel_i(ObgE_j\phi \rightarrow E_j\phi)$ [7] , où $E_j\phi$ signifie que $j$ a fait en sorte qu'on ait (est la cause de) $\phi$. Bien que ce ne soit pas explicité dans les formules, pour Jones les propriétés attribuées à $j$ ne sont pas satisfaites qu'à l'instant présent, mais toujours dans le futur.

Une autre différence importante porte sur le fait que dans l'approche de Castelfranchi et al. $i$ croit que $j$ satisfera son but en réalisant une action bien déterminée $\alpha$, tandis que pour Jones cette action n'est pas nécessairement constitutive de la confiance. On peut résumer ces deux positions en disant que dans la première $i$ veut que son but soit atteint d'une certaine manière, alors que dans la seconde ce n'est que l'état à atteindre qui compte. Plus brièvement dans la seconde le but n'exprime que le "quoi", et la première exprime en plus le "comment". On peut formellement concilier ces deux approches en explicitant dans la formule $\phi$ qui définit le but, le fait que $j$ a réalisé l'action $\alpha$. Le but serait alors, par exemple, de la forme : $\phi \wedge Done_j\alpha$. Il resterait à préciser la modalité $Done_j\alpha$.

Par exemple, dans le cas du conducteur dont la roue est crevée, il se peut que $i$ veuille que la roue soit réparée sur place, ou, au contraire, que la voiture soit transportée dans un garage pour faire la réparation. A l'inverse, si $i$ approche d'un carrefour où il y a un "stop" pour les voitures qui viennent d'une route perpendiculaire, et que son but est qu'aucune autre voiture ne se trouve au carrefour pendant qu'il passe, il peut avoir confiance dans le fait qu'aucun conducteur $j$ venant de l'autre route n'essaiera de faire qu'il se trouve au milieu du carrefour. Et dans ce cas $i$ ne se préoccupe pas de l'action que pourrait faire $j$ pour se trouver au milieu du carrefour ; par exemple il se pourrait que $j$ soit arrêté et qu'il lui barre la route en faisant l'action de démarrer, ou bien que $j$ soit en train d'approcher du carrefour et qu'il s'abstienne de freiner pour arrêter sa voiture.

# 7 Conclusion

Partant d'une définition de la confiance en termes de croyance dans le fait que le but d'un agent sera satisfait, nous avons montré que ce type de confiance peut être justifié par la confiance dans certaines propriétés attribuées à d'autres agents : la capacité à obtenir un certain état des choses, la possibilité de transformer les intentions en actions, et enfin la possibilité d'adopter certaines intentions. Celle-ci peut être déterminée, soit par le respect des normes, soit par l'intérêt mutuel, soit par l'intérêt pour l'autre. Chacune de ces propriétés a sa duale selon qu'un agent attend d'un autre agent qu'il agisse pour satisfaire son but, ou qu'il s'abstienne d'agir pour l'empêcher de maintenir son but.

Grâce à la formalisation en logique modale on a pu mettre en évidence des distinctions entre des notions comme : obéissance et honnêteté, ou entre celles représentées par : $ActAltruis$ et $RefAltruis$. Ces dernières correspondent aux distinctions qu'ont mis en évidence J. Horty et N. Belnap dans [10] avec les modalités d'action du type "to see to it that" et "to refrain to see to it that".

D'autre part, en explicitant la dimension temporelle, nous avons mis en évidence la distinction entre la confiance qui est motivée par un événement particulier à l'instant présent, et celle qui est motivée par l'occurrence éventuelle de cet événement dans le futur.

Du point de vue de la formalisation, nous avons adopté un opérateur de conditionnelle pour lequel nous avons requis les propriétés les plus

7. Nous avons changé les notations pour les rendre homogènes avec le reste de l'article.

faibles, celles qui suffisent à déduire ce qu'il nous intéressait de déduire dans le cadre de cet article (essentiellement le fait que certains types de confiance permettent de déduire d'autres types de confiance), laissant ainsi la porte ouverte à l'adoption de schémas d'axiomes plus spécifiques si nécessaire. La formalisation de la signification de chaque opérateur modal n'a pas été développée, et nous nous sommes concentrés uniquement sur leurs interrelations.

De nombreux aspects restent à étudier et à formaliser au sujet de la confiance. Nous n'en mentionnerons que trois.

- Il faudrait distinguer des buts de natures différentes. Les buts épistémiques, quand le but de $i$ est de connaître une information (voir [5]), par exemple la situation financière d'une banque, ou une prévision relative à l'évolution du cours d'une action. Les buts "physiques", quand le but est d'atteindre une certaine situation physique de l'état du monde, par exemple qu'une roue soit réparée. Et enfin les buts déontiques, par exemple quand le but de $i$ est de créer une obligation en portant plainte auprès d'un magistrat.
- Dans l'analyse que nous avons faite dans cet article nous n'avons distingué que le cas où $i$ a, ou n'a pas, confiance. Il faudrait pouvoir distinguer des degrés de confiance. Ces nuances sont évidentes quand la justification de la confiance est de nature empirique et s'appuie sur des statistiques. Il reste beaucoup de points à étudier quand la justification n'est pas empirique et qu'on considère des degrés qualitatifs [15, 7].
- Enfin, dans le domaine de l'évolution de la confiance, il nous semble important d'aller au-delà des modifications de probabilités objectives qui peuvent résulter de l'acquisition de nouvelles observations. La confiance est une notion essentiellement subjective qui n'évolue pas toujours de façon rationnelle. Par exemple, un accident de la route très rare, mais spectaculaire, peut avoir un effet très fort sur la confiance même si la probabilité de cet accident reste inchangée. D'autre part, l'évolution de la confiance peut être influencée par des changements de nature normative et qui ne peuvent être pris en compte par des statistiques. Par exemple, le fait de réduire la vitesse maximum autorisée sur autoroute peut augmenter la confiance par rapport au risque d'accidents.

# Annexe

On suppose que l'axiomatique des connecteurs logiques est définie par le Calcul Classique des Propositions.

Pour l'opérateur de conditionnelle on adopte les règles d'inférences et schémas d'axiomes suivants.

(EQUIV) Si $\vdash \phi \leftrightarrow \phi'$ et $\vdash \psi \leftrightarrow \psi'$, alors $\vdash (\phi \Rightarrow \psi) \rightarrow (\phi' \Rightarrow \psi')$
(PROPG) $(\phi_1 \wedge \phi_2 \Rightarrow \phi_3) \rightarrow (\phi_1 \wedge \phi_2 \Rightarrow \phi_1 \wedge \phi_3)$
(TRANS) $(\phi_1 \Rightarrow \phi_2) \wedge (\phi_2 \Rightarrow \phi_3) \rightarrow (\phi_1 \Rightarrow \phi_3)$
(DIST) $(\phi_1 \wedge (\phi_1 \Rightarrow \phi_2) \wedge \psi) \Rightarrow (\phi_2 \wedge \psi)$
(MATIMP) $(\phi \Rightarrow \psi) \rightarrow (\phi \rightarrow \psi)$

Dans cette logique on peut facilement démontrer par induction à partir de (TRANS) et (DIST) que l'on a :

$$\phi_1 \wedge (\phi_1 \Rightarrow \phi_2) \wedge (\phi_2 \Rightarrow \phi_3) \wedge \ldots$$
$$\ldots \wedge (\phi_{n-1} \Rightarrow \phi_n) \Rightarrow \phi_n$$

# Références


[1] G. Boella, J. Hulstin, Y-H. Tan, and L. van der Torre. Transaction trust in normative multi agent systems. In *AAMAS Workshop on Trust in Agent Societies*, 2005.

[2] C. Castelfranchi and R. Falcone. Social trust : a cognitive approach. In C. Castelfranchi and Y-H. Tan, editors, *Trust and Deception in Virtual Societies*. Kluwer Academic Publisher, 2001.

[3] B. F. Chellas. *Modal Logic : An introduction*. Cambridge University Press, 1988.

[4] R. Demolombe. Syntactical Characterization of a Subset of Domain Independent Formulas. *Journal of ACM*, 39(1), 1982.

[5] R. Demolombe. To trust information sources : a proposal for a modal logical framework. In C. Castelfranchi and Y-H. Tan, editor, *Proc. of the Workshop on Deception, Fraud and Trust in Agent Societies*, 1998.

[6] R. Demolombe. Reasoning about trust : a formal logical framework. In C. Jensen, S. Poslad, and T. Dimitrakos, editors, *Trust management : Second International Conference iTrust (LNCS 2995)*. Springer Verlag, 2004.

[7] R. Demolombe and C-J. Liau. A logic of graded trust and belief fusion. In C. Castelfranci and R. Falcone, editors, *Proc. of 4th Workshop on Deception, Fraud and Trust*, 2001.

[8] R. Demolombe and V. Louis. Speech acts with institutional effects in agent societies. In L. Goble and J-J. Ch. Meyer, editors, *Deontic Logic and Artificial Normative Systems*. Springer, LNAI 4048, 2006.

[9] R. Falcone and C. Castelfranchi. Trust dynamics : How trust is influenced by direct experiences and by trust itself. In *Proceedings of the 3rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS-04)*, pages 740–747. New York, ACM, 2004.

[10] J.F. Horty and N. Belnap. The deliberative STIT : a study of action, omission, ability, and obligation. *Journal of Philosophical Logic*, 24 :583–644, 1995.

[11] A. J. Jones and M. Sergot. A formal characterisation of institutionalised power. *Journal of the Interest Group in Pure and Applied Logics*, 4(3), 1996.

[12] A.J.I. Jones. On the concept of trust. *Decision Support Systems*, 33, 2002.

[13] A.J.I. Jones and B.S. Firozabadi. On the characterisation of a trusting agent. Aspects of a formal approach. In C. Castelfranchi and Y-H. Tan, editors, *Trust and Deception in Virtual Societies*. Kluwer Academic Publisher, 2001.

[14] J. Ladrière. Les limites de la formalisation. In J. Piaget, editor, *Logique et Connaissance Scientifique*. Editions Gallimard, 1967.

[15] N. Laverny and J. Lang. From knowledge-based programs to graded belief-based programs part ii : Off-line reasoning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2005.

[16] E. Lorini and R. Demolombe. Trust and norms in the context of computer security : a logical formalization. In R. van der Meyden and L. van der Torre, editors, *Deontic Logic in Computer Science*. Springer, LNAI 5076, 2008.

[17] I. Pörn. Action Theory and Social Science. Some Formal Models. *Synthese Library*, 120, 1977.

[18] G. H. von Wright. *Norm and Action*. Routledge and Kegan, 1963.