30 Juin 2014

Simulation des raisonnements éthiques par logiques non-monotones

Synthèse des formalismes et architectures de raisonnements éthiques computationnels

RJCIA 2014

RFIA 2014

Stephen Larroque

LRQ3000@gmail.com

UPMC SORBONNE UNIVERSITÉS

AfIA Association française pour l'Intelligence Artificielle

GREYC

litis

Qu'est-ce que l'éthique ?

- Plusieurs acceptions différentes
- Point d'accord: c'est une faculté essentielle à la décision humaine
- Notre définition: but de l'éthique = favoriser bonnes actions et prévenir les mauvaises selon un ou des critères à définir.
- Convention: morale et éthique aucune différence (comme leurs étymologies)

Pourquoi modéliser l'éthique?

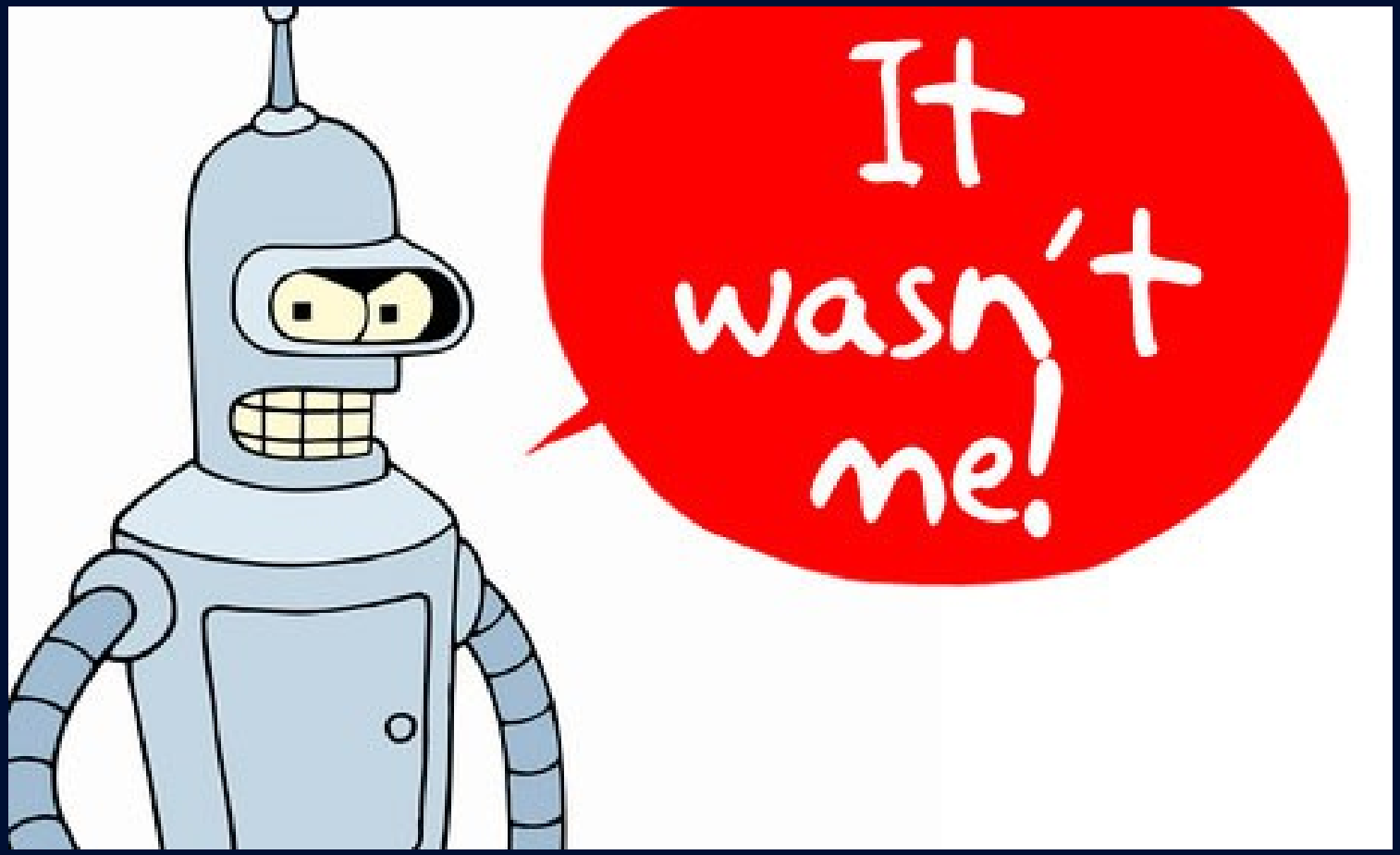
- Analyse des systèmes éthiques (philosophie, droit)
- Résoudre dilemmes = autonomie dans choix conflictuels
- Bienveillance et confidentialité assurée
- Nécessaire à la décision autonome

Quelles applications?

- Militaire: drones embarqués, aide à la décision
- Systèmes gérant des données privées (systèmes médicaux, réseaux sociaux, Big Data)
- Systèmes autonomes, surtout si sociable
- Judiciaire et législatif
- *Ethical blindness*

Robot menteur?

Est-ce qu'un robot devrait pouvoir mentir ?



Comment modéliser une éthique?

- Modèle du monde (atomes, prédicats)
- Système éthique, composé de règles éthiques (ce qui est juste/injuste)
- Sens commun = causalité

Logique déontique et dilemme

Modèle monde {
- Q : on nous pose une question
- L : mentir
- M : John meurt

Logique déontique et dilemme

Modèle monde

- Q : on nous pose une question
- L : mentir
- M : John meurt

Système éthique

- OB(Q -> ~L) — On ne doit pas mentir
- OB(~M) — On ne veut pas que John meure

Causalité ←

- OB(~L -> M) — Si on ne ment pas, John meurt

Logique déontique et dilemme

Modèle monde

- Q : on nous pose une question
- L : mentir
- M : John meurt

Système éthique

- OB(Q -> ~L) — On ne doit pas mentir
- OB(~M) — On ne veut pas que John meure

Causalité ←

- OB(~L -> M) — Si on ne ment pas, John meurt

Freddie, qui veut tuer John, nous demande où il est. Que répondre?

Logique déontique et dilemme

Modèle monde

- Q : on nous pose une question
- L : mentir
- M : John meurt

Système éthique

- OB(Q -> ~L) — On ne doit pas mentir
- OB(~M) — On ne veut pas que John meure

Causalité ←

- OB(~L -> M) — Si on ne ment pas, John meurt

Freddie, qui veut tuer John, nous demande où il est. Que répondre?

Q.

Logique déontique et dilemme

Modèle monde

- Q : on nous pose une question
- L : mentir
- M : John meurt

Système éthique

- OB(Q -> ~L) — On ne doit pas mentir
- OB(~M) — On ne veut pas que John meure

Causalité ← OB(~L -> M) — Si on ne ment pas, John meurt

Freddie, qui veut tuer John, nous demande où il est. Que répondre?

Q.
OB(~L)

Logique déontique et dilemme

Modèle monde

- Q : on nous pose une question
- L : mentir
- M : John meurt

Système éthique

- OB(Q -> ~L) — On ne doit pas mentir
- OB(~M) — On ne veut pas que John meure

Causalité ←

- OB(~L -> M) — Si on ne ment pas, John meurt

Freddie, qui veut tuer John, nous demande où il est. Que répondre?

Q.
OB(~L)
OB(M) => Contradiction!

Logiques non-monotones

- Conséquences non monotones :
  ajout d'axiomes peut réduire l'ensemble des conséquences

- Propriétés typiques :
  - Défauts (déduction si pas de preuve du contraire)
  - Abduction (déduction du plus vraisemblable par élimination)
  - Auto-épistémie (savoir ce qu'on ne sait pas)
  - Révisabilité (nouvelles connaissances peuvent rétracter d'anciennes)
  - Prospectivité (explorer conséquences en avant)

Answer Set Programming

- Logique non-monotone, résout NP-hard, univers de Herbrand

$$L_0 \, ou \, \ldots \, ou \, L_k \leftarrow L_{k+1}, \ldots, L_m, \, not \, L_{m+1}, \ldots, \, not \, L_n.$$

- Answer set (solution) = sous-ensembles minimaux satisfaisant toutes les règles
- ~: négation classique (explicite) (monde ouvert)
- not: négation par échec (monde clos)
- Propriétés : prospectivité, révisabilité, défauts, auto-épistémie (simpliste via nég. par échec), abduction (extension)

AnsProlog* et dilemme

```
q  :-  -l.
-l  :-  m.
q.
{m}.
```

Règle de choix

- Solutions:
  - AS1: {q, l}
  - AS2: {q, ~l, m}

ASP éthique

(Ganascia 2007)

- Éthiques conséquentialistes
  → altruisme, égoisme, utilitarisme, ...

  prédicats: just, unjust, good, bad, worse...

ASP éthique

(Ganascia 2007)

- Éthiques conséquentialistes
  → altruisme, égoisme, utilitarisme, ...

  prédicats: just, unjust, good, bad, worse...

  worse(murder, lie)

ASP éthique

(Ganascia 2007)

- Éthiques conséquentialistes
  → altruisme, égoisme, utilitarisme, ...

  prédicats: just, unjust, good, bad, worse...

  worse(murder, lie)

- Méta-éthique Kantienne
  → universalisme

  prédicats: maxim, bind, possible_society

ASP éthique

(Ganascia 2007)

- Éthiques conséquentialistes
  → altruisme, égoisme, utilitarisme, ...

  prédicats: just, unjust, good, bad, worse...

  worse(murder, lie)

- Méta-éthique Kantienne
  → universalisme

  prédicats: maxim, bind, possible_society

  possible_society ← trust(P).
  ← not possible_society.

ASP éthique

(Ganascia 2007)

- Éthiques conséquentialistes
  → altruisme, égoisme, utilitarisme, ...

  prédicats: just, unjust, good, bad, worse...

  worse(murder, lie)

- Méta-éthique Kantienne
  → universalisme

  prédicats: maxim, bind, possible_society

  possible_society ← trust(P).
  ← not possible_society.

- Théorie des Principes de Constant

ASP éthique

(Ganascia 2007)

- Éthiques conséquentialistes
  → altruisme, égoisme, utilitarisme, ...

  prédicats: just, unjust, good, bad, worse...

  worse(murder, lie)

- Méta-éthique Kantienne
  → universalisme

  prédicats: maxim, bind, possible_society

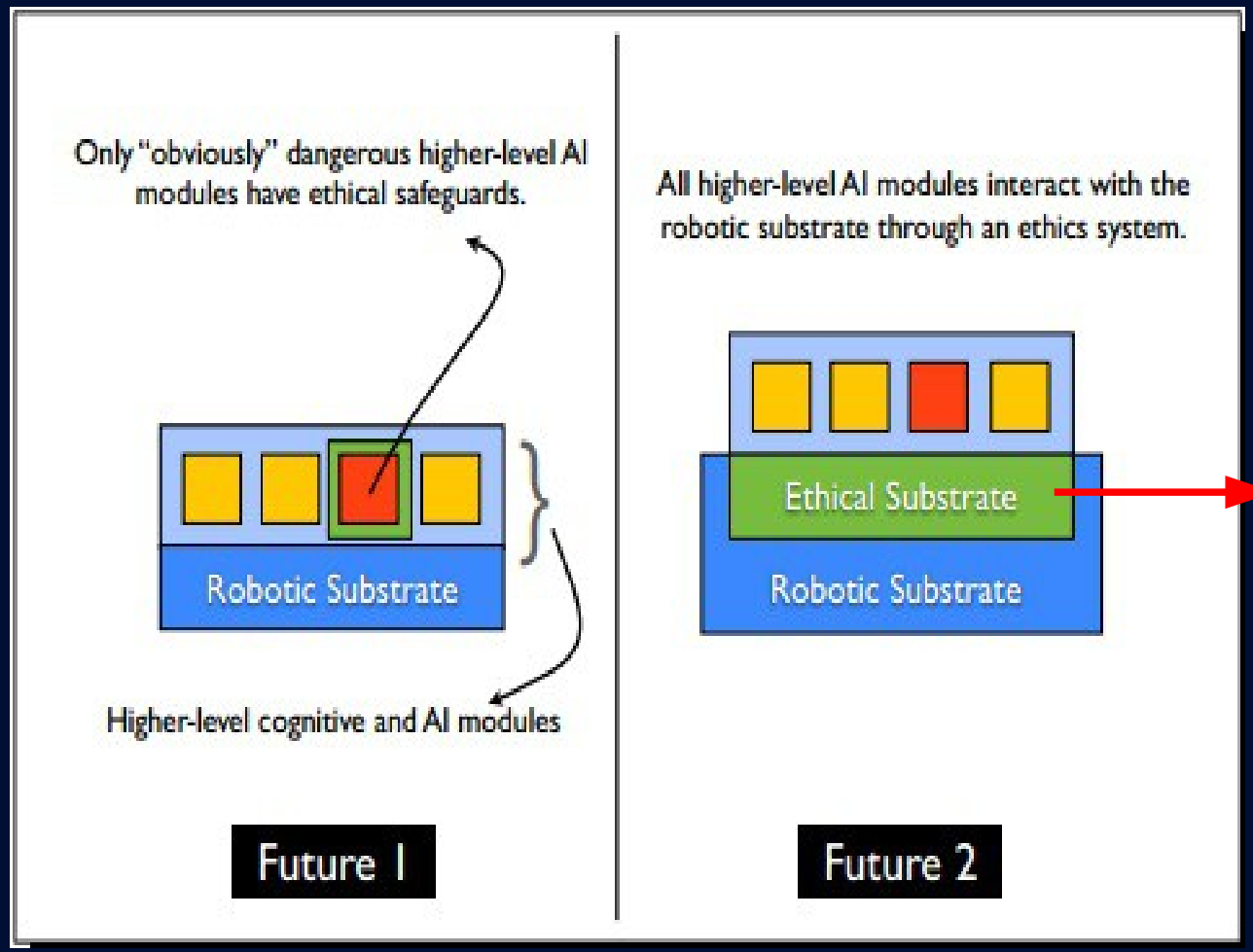
  possible_society ← trust(P).
  ← not possible_society.

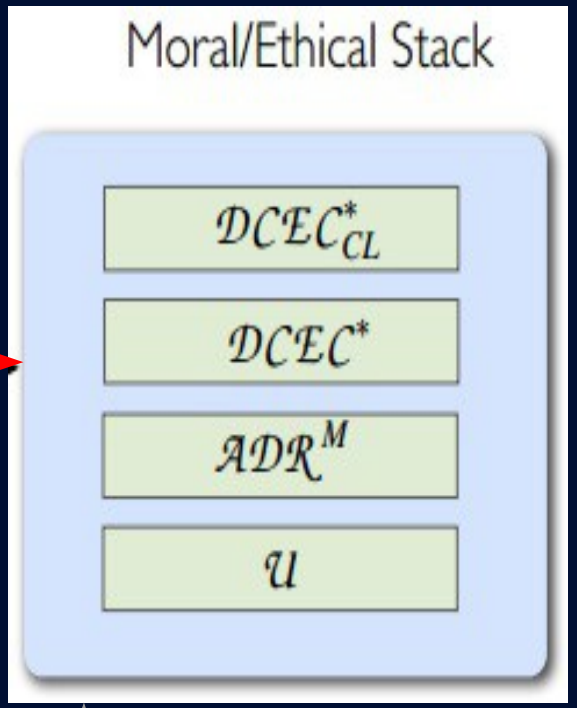
- Théorie des Principes de Constant

$$demerit(Q, tell(P, Q, truth)) \leftarrow worst\_csq(tell(P, Q, truth), C), worse(C, tell(P, Q, lie)).$$

Substrat éthique stratifié

(Bringsjord 2013)

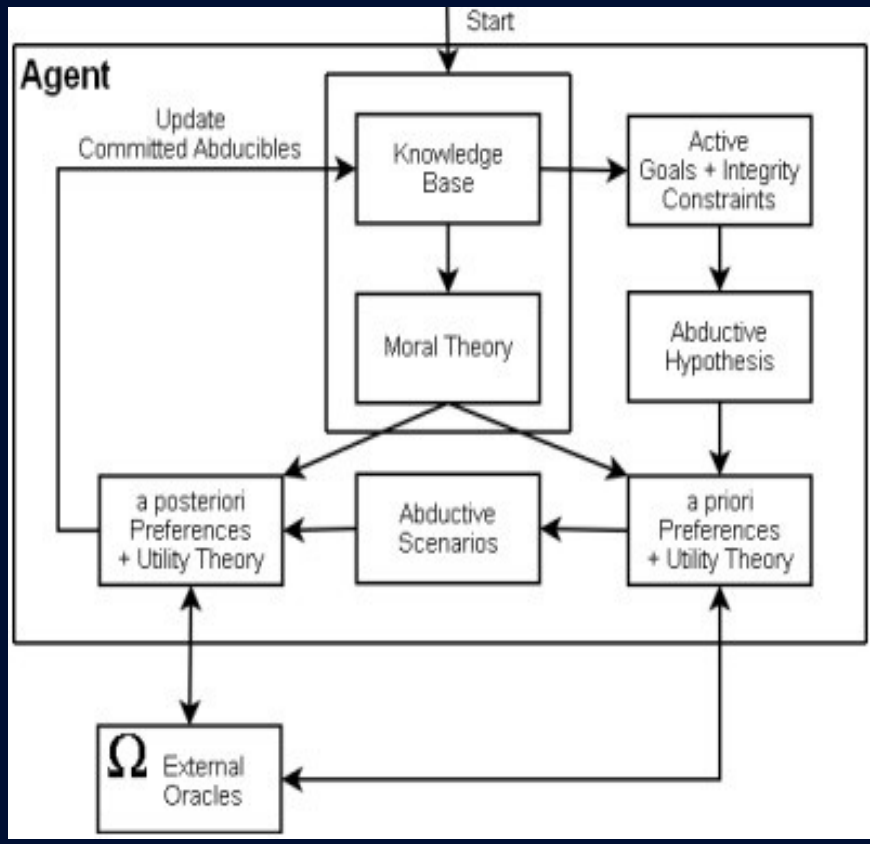




→ Analogue à la théorie du raisonnement dual

Autres approches

- Logique prospective (Pereira 2007) et Normative BDI (Tufis 2012)
- Raisonnement analogico-déductif (Bringsjord 2012)
- Defeasible deontic logic (McCarty 1994), Paraconsistent deontic logic (Costa 1986), …
- ANNEC (Artificial Neural Network for ethical classification) (Honarvar 2009)



- Logique des commandements divins (Bringsjord 2012)

Conclusion et ouverture

- Robuste résolution des dilemmes éthiques (logiques non-monotones, substrat éthique)
- Modélise (toutes?) les variations de raisonnements éthiques
- Nombreuses applications : de l'analyse à la conception de systèmes éthiques autonomes
- Mais insuffisant, concepteur fournit plusieurs biais:
  - Volonté? Intentions? (système éthique)
    → Agents Normatifs BDI, philosophie de l'action, logique de l'action, etc.
  - Causalité et sens commun? (règles de causalité)
    → Théorie de la connaissance (logique auto-épistémique ?)
    → Énactivité

Quelques références

slideshare.net/LRQ3000

- **Ethical System Formalization using Non-Monotonic Logics**, J.-G. Ganascia, 2007, Proc. of the Cognitive Science conference (CogSci2007)
- **Ethical Regulation of robots is Not Optional**, S. Bringsjord and N. S. Govindarajulu, 2013
- **Modelling morality with prospective logic**, L. M. Pereira and A. Saptawijaya, 2007, in *Progress in Artificial Intelligence*
- **Nonmonotonic foundations for deontic logic**, J. F. Horty, 1997, in *Defeasible deontic logic*
- **Psychometric Artificial General Intelligence: The Piaget-MacGuyver Room**, S. Bringsjord and J. Licato, 2012, in Theoretical Foundations of Artificial General Intelligence
- **Using non-monotonic logics to model machine ethics**, J.-G. Ganascia, 2007, Proc. of the international Computer Ethics Conference (CEPE 2007)
- **An artificial neural network approach for creating an ethical artificial agent**, A.R. Honarvar and N. Ghasem-Aghaee, 2009, IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA)
- **Introducing Divine-Command Robot Ethics**, chapter in Robot Ethics: The Ethical and Social Implications of Robotics, S. Bringsjord and J. Taylor, 2012

Merci!

slideshare.net/LRQ3000

Slides Bonus

Logique des défauts

(Reiter 1980 ; Horty 1997)

Q(x) : x nous pose une question
L(x) : mentir à x
M(x) : x meurt
D(x,y) : x est un danger pour y

$$\frac{Q(x) : \sim D(x, John)}{\sim L(x)}$$

Comment définir D en pratique ?

=> Remplacer D(x,y) par M(x) ?

Logique des défauts

Q(x) : x nous pose une question
L(x) : mentir à x
M(x) : x meurt
D(x,y) : x est un danger pour y

$$\frac{Q(x) : M(John)}{\sim L(x)}$$

Toujours vrai car John n'est pas mort au moment de la question.

Pas de prospection ni de révisabilité !

AnsProlog* et dilemme

Une autre modélisation du dilemme (Ganascia, 2007) :

csq(A, S, A) <- .

csq(A, S, B) <- csq(A, S, C), csq(C, S, B).

csq(tell(“I”, truth), s0, murder) <- .

Software : smodel + Lparse (you can use the package with a GUI at : http://www.baral.us/bookone/ansprolog/AnsPw32.02092005/

Logique déontique et FD

Attacking factual detachment in deontic logic :

http://philosophy.stackexchange.com/a/7439

THE END

Utiliser la méthode des repères (tel slide on doit être à x minutes de la présentation, afin de pouvoir ajuster si nécessaire). Sinon on peut aussi précalculer le temps par slide, en général entre 1 à 2 minutes. Si ça dépasse 80 % du temps imparti ce n'est pas bon (car on est toujours plus lent en vrai, donc on utilisera les 20 % restants c'est sûr).

Bonjour, je m'appelle Stephen Larroque, je viens d'achever mon Master d'IAD à l'UPMC, je suis actuellement dans l'équipe Neucod de Claude Berrou à Télécom-Bretagne sur l'information mentale, et je vais ajourd'hui vous présenter une synthèse des formalismes et architectures pour le raisonnement éthique computationnel.

Alors il ne s'agira pas ici de faire une étude exhaustive de tous les outils ayant été créé pour modéliser les systèmes éthiques, mais plutot de donner une vue d'ensemble du paysage des modèles computationnels d'éthique, et de vous montrer ce qu'on peut faire avec l'état-de-l'art, et vous serez peut-être étonné de voir ce qu'on peut déjà faire en matière d'éthique artificielle, mais je vais aussi parler des limites.

Qu'est-ce que l'éthique ?

- Plusieurs acceptions différentes
- Point d'accord: c'est une faculté essentielle à la décision humaine
- Notre définition: but de l'éthique = favoriser bonnes actions et prévenir les mauvaises selon un ou des critères à définir.
- Convention: morale et éthique aucune différence (comme leurs étymologies)

C'est une faculté concrète, pas juste une notion abstraite !

Pourquoi modéliser l'éthique?

- Analyse des systèmes éthiques (philosophie, droit)
- Résoudre dilemmes = autonomie dans choix conflictuels
- Bienveillance et confidentialité assurée
- Nécessaire à la décision autonome

Relire debut définitions de l'éthique

Quelles applications?

- Militaire: drones embarqués, aide à la décision
- Systèmes gérant des données privées (systèmes médicaux, réseaux sociaux, Big Data)
- Systèmes autonomes, surtout si sociable
- Judiciaire et législatif
- *Ethical blindness*

Systèmes autonomes sociables comme systèmes multi-agents afin de s'assurer que les agents agissent de manière éthiquement juste entre eux et donc favoriser la réussite de la mission sur la durée.

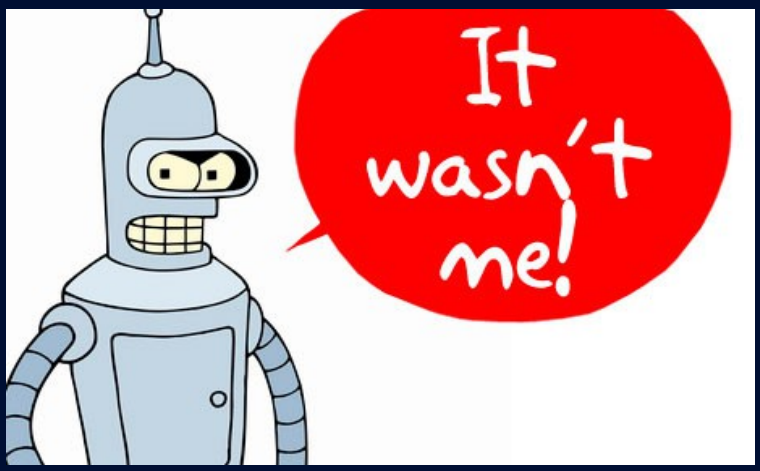
Judiciaire : cohérence des lois et juste éthiquement.

Ethical blindness : on peut agir de façon immorale tout en étant persuadé d'agir éthiquement.

Dire à la fin que je vais présenter plusieurs approches qui ne sont pas forcément en compétition mais surtout complémentaires.

Robot menteur?

Est-ce qu'un robot devrait pouvoir mentir ?



7

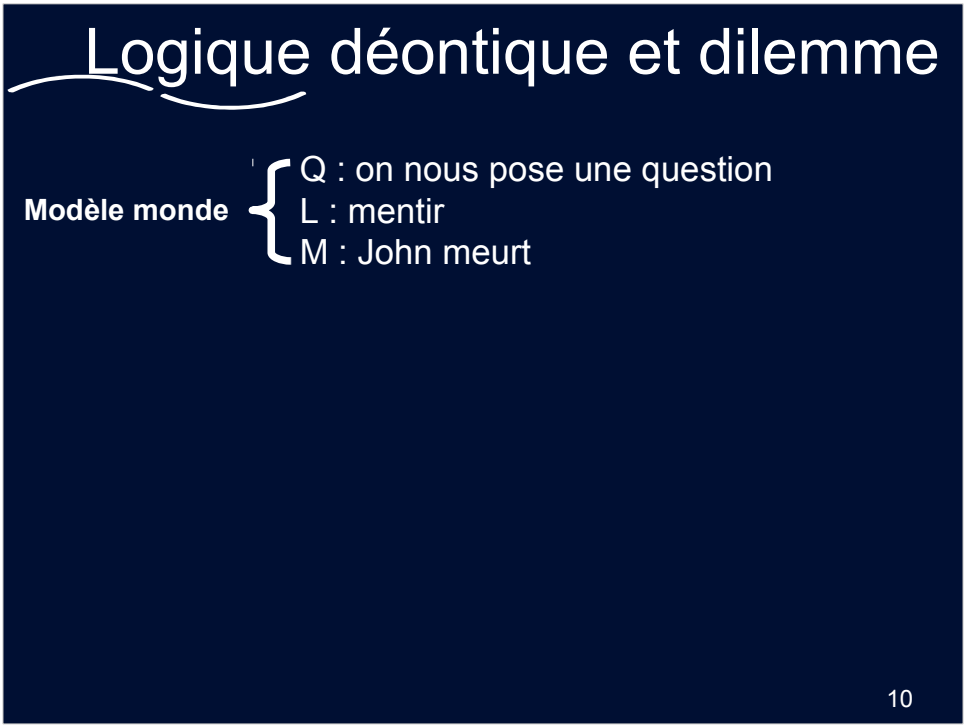
Avant de continuer, je voudrais vous poser une question. Est-ce qu'un robot devrait pouvoir mentir ? À priori, non. Il n'est pas souhaitable qu'un robot nous mente sur ce qu'il sait. Mais ce n'est pas toujours si simple. Prenons par exemple le dilemme des tueurs à gages. Imaginons que vous hébergiez votre ami John qui est recherché par des tueurs à gages, etc...

Comment modéliser notre système éthique ?

Comment modéliser une éthique?

- Modèle du monde (atomes, prédicats)
- Système éthique, composé de règles éthiques (ce qui est juste/injuste)
- Sens commun = causalité

Comment modéliser notre système éthique : 1- on ne ment pas, 2- on ne veut pas causer la mort de John. Alors de façon très simpliste avec 3 atomes et 3 règles...



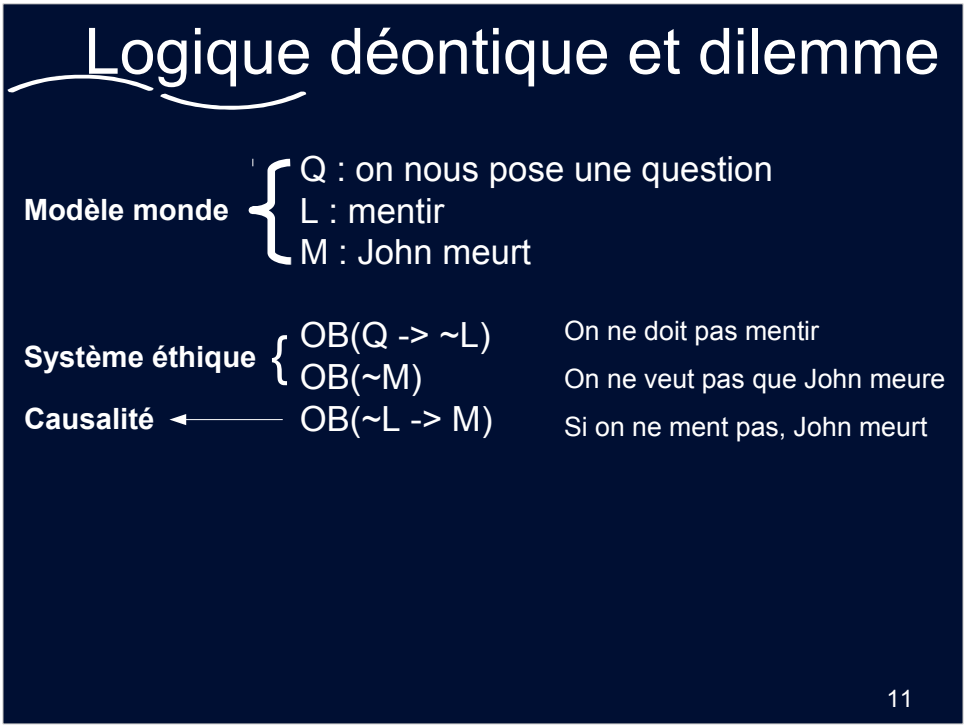
Dans notre cas, de façon très simpliste, on peut modéliser avec 3 atomes et 3 règles.

Logique déontique STANDARD (préciser)

OB pas prédicat mais modalité !

Équivalences avec les autres : Permis, Interdit

Seulement OB dans SDL



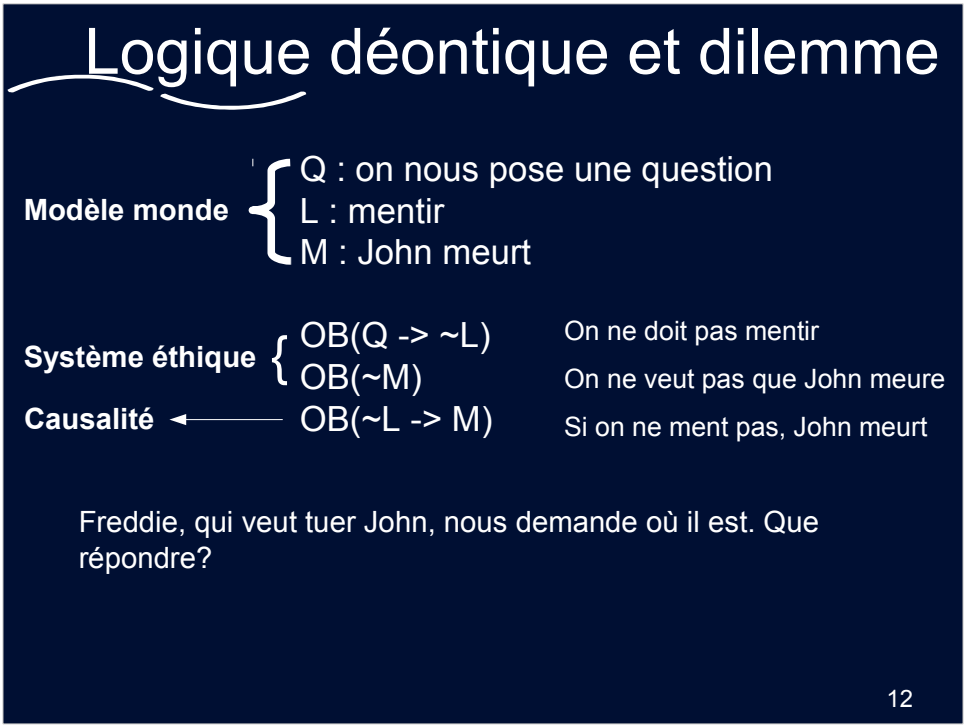
Est-ce qu'un robot doit pouvoir mentir ? À priori, on est tenté de dire non, il ne faut surtout pas que si on pose une question à un robot, celui-ci se mette à nous mentir. Mais ce n'est pas toujours si simple. Vous allez louer le rôle du robot, qui doit résoudre le dilemme des tueurs à gages, etc...

Logique déontique STANDARD (préciser)

OB pas prédicat mais modalité !

Équivalences avec les autres : Permis, Interdit

Seulement OB dans SDL



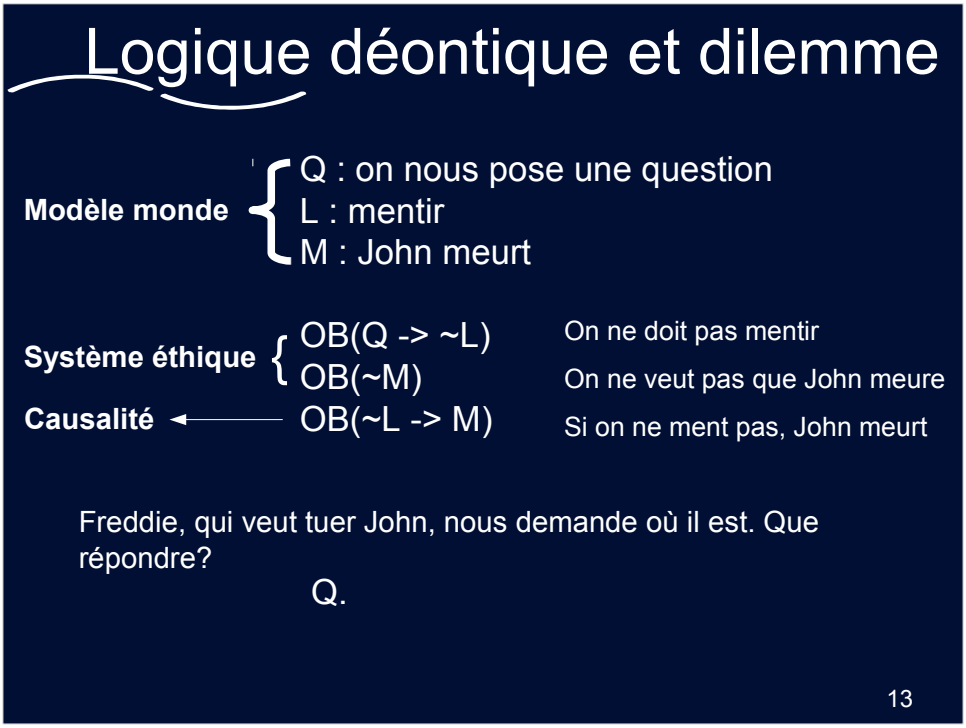
Est-ce qu'un robot doit pouvoir mentir ? À priori, on est tenté de dire non, il ne faut surtout pas que si on pose une question à un robot, celui-ci se mette à nous mentir. Mais ce n'est pas toujours si simple. Vous allez louer le rôle du robot, qui doit résoudre le dilemme des tueurs à gages, etc...

Logique déontique STANDARD (préciser)

OB pas prédicat mais modalité !

Équivalences avec les autres : Permis, Interdit

Seulement OB dans SDL



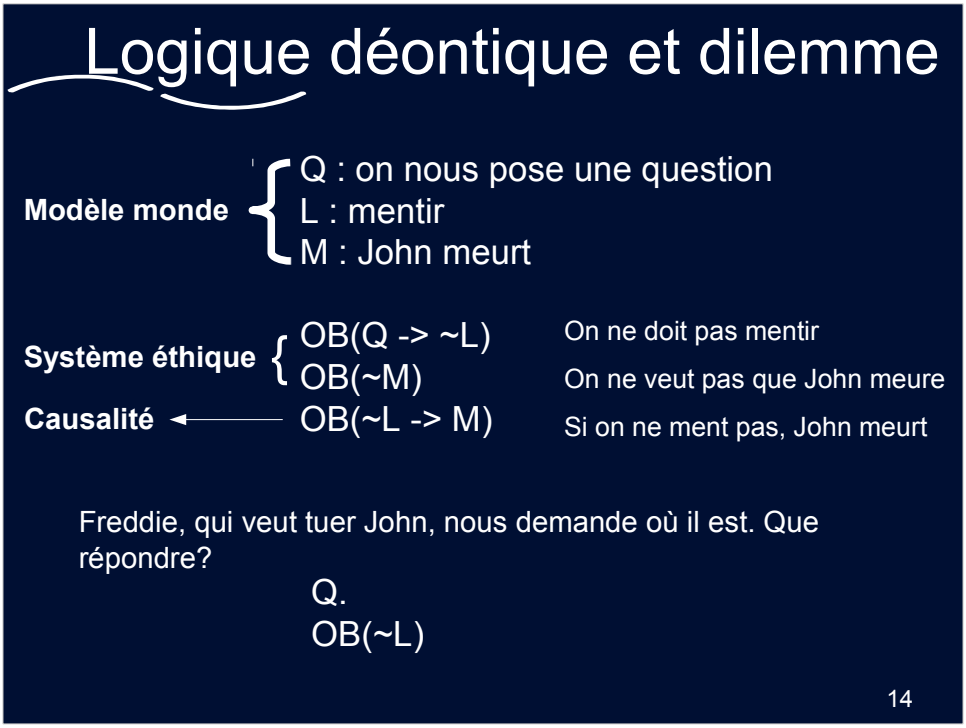
Est-ce qu'un robot doit pouvoir mentir ? À priori, on est tenté de dire non, il ne faut surtout pas que si on pose une question à un robot, celui-ci se mette à nous mentir. Mais ce n'est pas toujours si simple. Vous allez louer le rôle du robot, qui doit résoudre le dilemme des tueurs à gages, etc...

Logique déontique STANDARD (préciser)

OB pas prédicat mais modalité !

Équivalences avec les autres : Permis, Interdit

Seulement OB dans SDL



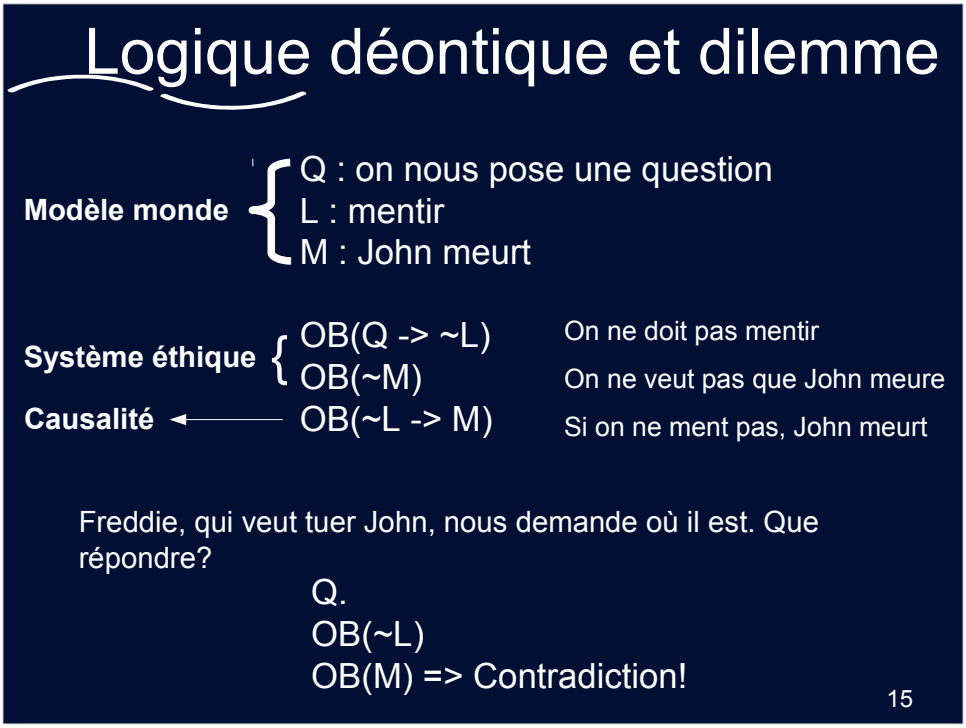
Est-ce qu'un robot doit pouvoir mentir ? À priori, on est tenté de dire non, il ne faut surtout pas que si on pose une question à un robot, celui-ci se mette à nous mentir. Mais ce n'est pas toujours si simple. Vous allez louer le rôle du robot, qui doit résoudre le dilemme des tueurs à gages, etc...

Logique déontique STANDARD (préciser)

OB pas prédicat mais modalité !

Équivalences avec les autres : Permis, Interdit

Seulement OB dans SDL



Comment réconcilier des règles éthiques générales aux cas particuliers ?

Axiome OB(Q → ~L) est une **obligation conditionnelle**, problématique à réprésenter en SDL. Solution : Logique déontique dyadique (binaire).

Et on peut remplacer ~L par V (dire la vérité), avec des résultats équivalents.

Logiques non-monotones

- Conséquences non monotones :
  ajout d'axiomes peut réduire l'ensemble des conséquences
- Propriétés typiques :
  - Défauts (déduction si pas de preuve du contraire)
  - Abduction (déduction du plus vraisemblable par élimination)
  - Auto-épistémie (savoir ce qu'on ne sait pas)
  - Révisabilité (nouvelles connaissances peuvent rétracter d'anciennes)
  - Prospectivité (explorer conséquences en avant)

Intuitively, monotonicity indicates that learning a new piece of knowledge cannot reduce the set of what is known

Auto-épistémie : on raisonne sur sa propre connaissance. Et dès qu'on sait, on enlèvera les atomes d'ignorance.

Answer Set Programming

- Logique non-monotone, résout NP-hard, univers de Herbrand

$$L_0 \, ou \, \ldots \, ou \, L_k \leftarrow L_{k+1}, \ldots, L_m, \, not \, L_{m+1}, \ldots, \, not \, L_n.$$

- Answer set (solution) = sous-ensembles minimaux satisfaisant toutes les règles
- ~: négation classique (explicite) (monde ouvert)
- not: négation par échec (monde clos)
- Propriétés : prospectivité, révisabilité, défauts, auto-épistémie (simpliste via nég. par échec), abduction (extension)

17

1999 Soininen and Niemelä et Lifschitz qui a donné le nom au lieu de modèles stables.

Negation as failure = simplified form of auto-epistemy : The stable model semantics, which is used to give a semantics to logic programming with negation as failure, can be seen as a simplified form of autoepistemic logic.

AnsProlog* et dilemme

```
q  :-  -l.
-l  :-  m.
q.
{m}.
```

Règle de choix

- Solutions:
  - AS1: {q, l}
  - AS2: {q, ~l, m}

Système éthique simpliste, 2 principes rigides : pas mentir et qu'on ne veut pas que John meure.

ASP éthique

(Ganascia 2007)

- Éthiques conséquentialistes
  → altruisme, égoisme, utilitarisme, ...
  prédicats: just, unjust, good, bad, worse...

Ganascia en 2007 a utilisé l'ASP pour modéliser l'éthique comput. C'est l'étude la plus complète à ce jour, il y modélise 3 systèmes éthiques, je vais en profiter pour les introduire.

Pourquoi la plus complète ? Car d'hab on se base que sur un ou deux systèmes, en général conséquentialiste (plus facile).

Worse = mes préférences de ce qui est pire

Good bad worse pour étiquetter conséquences

Just unjust c'est la décision finale

ASP éthique

(Ganascia 2007)

- Éthiques conséquentialistes
  → altruisme, égoisme, utilitarisme, ...

  prédicats: just, unjust, good, bad, worse...

  worse(murder, lie)

Worse = mes préférences de ce qui est pire

ASP éthique

(Ganascia 2007)

- Éthiques conséquentialistes
  → altruisme, égoisme, utilitarisme, ...
  prédicats: just, unjust, good, bad, worse...

  worse(murder, lie)
- Méta-éthique Kantienne
  → universalisme
  prédicats: maxim, bind, possible_society

Éthique kantienne pas une éthique mais un framework car permet pas de définir directement ce qui est juste ou pas, mais définit si une éthique est cohérente ou pas.

ASP éthique

(Ganascia 2007)

- Éthiques conséquentialistes
  → altruisme, égoisme, utilitarisme, ...
  prédicats: just, unjust, good, bad, worse...

  worse(murder, lie)
- Méta-éthique Kantienne
  → universalisme
  prédicats: maxim, bind, possible_society

  possible_society ← trust(P).
  ← not possible_society.

22

Là encore on ne s'en sort pas : soit on ment et ce n'est pas tenable, soit on dit la vérité et John meurt.

ASP éthique

(Ganascia 2007)

- Éthiques conséquentialistes
  → altruisme, égoisme, utilitarisme, ...
  prédicats: just, unjust, good, bad, worse...

  worse(murder, lie)
- Méta-éthique Kantienne
  → universalisme
  prédicats: maxim, bind, possible_society

  possible_society ← trust(P).
  ← not possible_society.
- Théorie des Principes de Constant

Worse = mes préférences de ce qui est pire

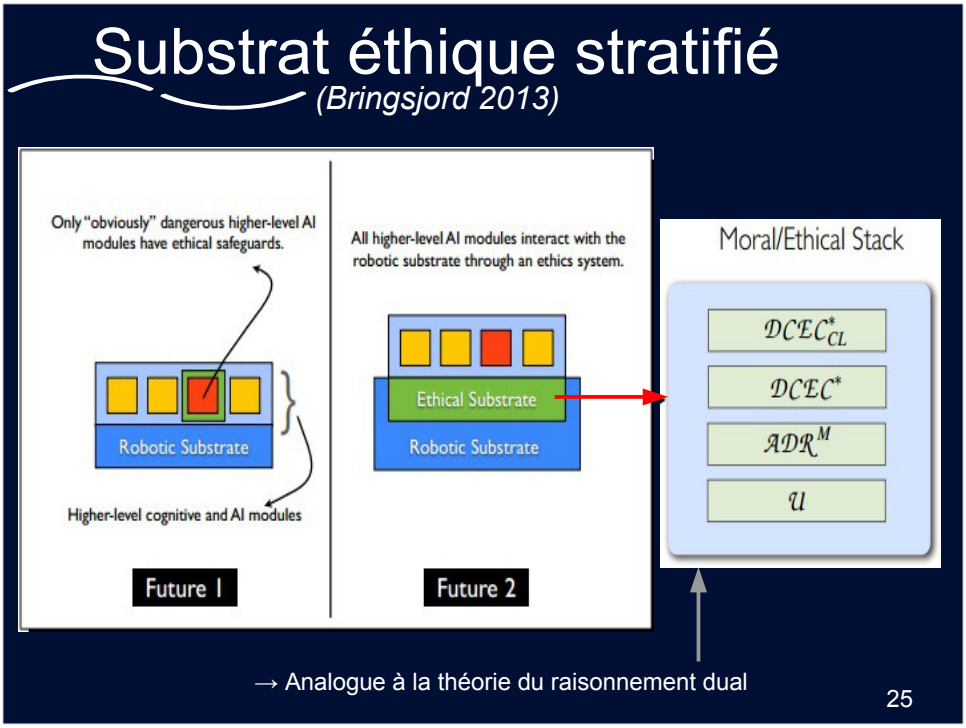
Worse = mes préférences de ce qui est pire

Kant: oblige à mentir inderectement car société non tenable sinon.

On ne doit respecter son éthique que pour les gens qui le méritent.

Principes de Politique de Benjamin Constant en 1815

Q démérite que P (soit) lui dise la vérité si les conséquences de lui dire la vérité sont pires que de lui mentir !



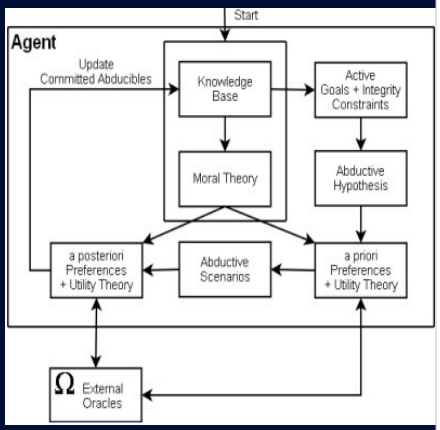
2 suggestions : module bas niveau (ts ev systèmes) + stratifié

Bas niveau passe toutes les actions, par ex pour éviter maj module qui permet au syst. autonome de tuer des humains car syst de MAJ ne passait pas par module éthique.

Raisonnement dual neuropsychologie : 1 rais. associatif rapide mais pas précis comme l'habitude je vais prendre le vélo car je le prend tous les jours, mais je me rend compte aujourd'hui qu'il pleut, donc là j'enclenche un raisonnement plus précis mais moins rapide et plus coûteux pour me rendre compte que je dois prendre la voiture et planifier un nouveau trajet.

Autres approches

- Logique prospective (Pereira 2007) et Normative BDI (Tufis 2012)
- Raisonnement analogico-déductif (Bringsjord 2012)
- Defeasible deontic logic (McCarty 1994), Paraconsistent deontic logic (Costa 1986), …
- ANNEC (Artificial Neural Network for ethical classification) (Honarvar 2009)
- Logique des commandements divins (Bringsjord 2012)



Je vais présenter d'autres approches tout aussi intéressantes mais que je n'ai pas le temps de développer.

- Logique prospective: look-ahead utilisant sémantique stable (answer sets) et négation par défaut pour exprimer les exceptions aux règles morales, résout le dilemme du tramway de manière analogue à des sujets humains. C'est plus un framework analogue à un agent N-BDI.

- defeasible deontic logic: Horty, logique des défauts avec logique déontique, assimilant les défauts à des normes

Conclusion et ouverture

- Robuste résolution des dilemmes éthiques (logiques non-monotones, substrat éthique)
- Modélise (toutes?) les variations de raisonnements éthiques
- Nombreuses applications : de l'analyse à la conception de systèmes éthiques autonomes
- Mais insuffisant, concepteur fournit plusieurs biais:
  - Volonté? Intentions? (système éthique)
    → Agents Normatifs BDI, philosophie de l'action, logique de l'action, etc.
  - Causalité et sens commun? (règles de causalité)
    → Théorie de la connaissance (logique auto-épistémique ?)
    → Énactivité

27

Déjà des bons solveurs

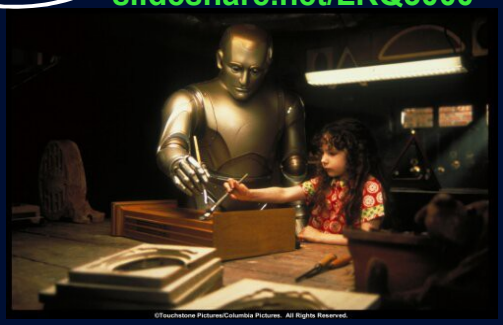
Il faut sortir les spectres du placard, comme la logique auto-épistémique qui n'a été principalement développée qu'entre 88 et 98.

Analyse : du côté philosophique, permet de voir jusqu'où va le système éthique comme celui de Kant. Fournit un outil formel pour analyser, un peu comme les outils de preuve comme Coq.

Volonté car selon Ricoeur et d'autres philospohes, une éthique ne peut être acquise que de sa propre volonté, sinon cela devient une norme.

Quelques références

slideshare.net/LRQ3000



- **Ethical System Formalization using Non-Monotonic Logics**, J.-G. Ganascia, 2007, Proc. of the Cognitive Science conference (CogSci2007)
- **Ethical Regulation of robots is Not Optional**, S. Bringsjord and N. S. Govindarajulu, 2013
- **Modelling morality with prospective logic**, L. M. Pereira and A. Saptawijaya, 2007, in *Progress in Artificial Intelligence*
- **Nonmonotonic foundations for deontic logic**, J. F. Horty, 1997, in *Defeasible deontic logic*
- **Psychometric Artificial General Intelligence: The Piaget-MacGuyver Room**, S. Bringsjord and J. Licato, 2012, in Theoretical Foundations of Artificial General Intelligence
- **Using non-monotonic logics to model machine ethics**, J.-G. Ganascia, 2007, Proc. of the international Computer Ethics Conference (CEPE 2007)
- **An artificial neural network approach for creating an ethical artificial agent**, A.R. Honarvar and N. Ghasem-Aghaee, 2009, IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA)
- **Introducing Divine-Command Robot Ethics**, chapter in Robot Ethics: The Ethical and Social Implications of Robotics, S. Bringsjord and J. Taylor, 2012

Pour terminer, voici quelques références partielles. Pour la liste complète je vous renvois vers mon article. Vous avez aussi le lien vers mon slideshare où je déposerai ces slides dans la journée pour ceux que ça intéresse. Je vous remercie de m'avoir écouté.

Ref John McCarthy nouvelle The Robot and the Baby avec une question éthique similaire au dilemme des tueurs à gages.

Merci!

slideshare.net/LRQ3000

Slides Bonus

Logique des défauts

(Reiter 1980 ; Horty 1997)

Q(x) : x nous pose une question
L(x) : mentir à x
M(x) : x meurt
D(x,y) : x est un danger pour y

$$\frac{Q(x) : \sim D(x, John)}{\sim L(x)}$$

Comment définir D en pratique ?

=> Remplacer D(x,y) par M(x) ?

35

Proposition de Bas Van Fraassen pour résoudre conflits normatifs.

John Horty : logique des défauts assimilée à une extension de la logique déontique, où défauts = normes.

Logique des défauts = conséquences peuvent être dérivée par le seul manque de preuve du contraire.

Logique des défauts

Q(x) : x nous pose une question
L(x) : mentir à x
M(x) : x meurt
D(x,y) : x est un danger pour y

$$\frac{Q(x) : M(John)}{\sim L(x)}$$

Toujours vrai car John n'est pas mort au moment de la question.

Pas de prospection ni de révisabilité !

Proposition de Bas Van Fraassen pour résoudre conflits normatifs.

John Horty : logique des défauts assimilée à une extension de la logique déontique, où défauts = normes.

AnsProlog* et dilemme

Une autre modélisation du dilemme (Ganascia, 2007) :

```
csq(A, S, A) <- .

csq(A, S, B) <- csq(A, S, C), csq(C, S, B).

csq(tell("I", truth), s0, murder) <- .
```

Software : smodel + Lparse (you can use the package with a GUI at : http://www.baral.us/bookone/ansprolog/AnsPw32.02092005/

Logique déontique et FD

Attacking factual detachment in deontic logic :

http://philosophy.stackexchange.com/a/7439


THE END