Ethics and Autonomous Agents

www.gregory.bonnet.free.fr
ethicaa.org

Grégory Bonnet

Université de Caen Normandie – GREYC

5 juillet 2017

Introduction
Problématique générale

Des humains et des agents interagissent au sein de systèmes

- logiciels : informatique ubiquitaire, aide à la décision, diagnostic
- robots : compagnions, véhicules autonomes, drones
- humains : opérateurs, utilisateurs

Comment réguler ces systèmes lorsque :

- les comportements attendus ne sont pas uniquement définis par des lois ?
- les comportements attendus peuvent l'être en vertu de valeurs subjectives ?
- une pluralité de valeurs doivent être respectées ?

Un projet transdisciplinaire

- intelligence artificielle (Armines-Fayol, Greyc, LIP6, Onera)
- ingénierie des connaissances (Ardans)
- éthique et sciences humaines et sociales (IMT)

Objectifs scientifiques

Concevoir des systèmes d'agents artificiels capables de :

- se représenter et raisonner sur l'éthique (normes, valeurs, principes)
- exprimer des compromis entre éthique et intérêts individuels
- justifier leurs décisions
- gérer des conflits éthiques entre agents

Concepts principaux
Morale

Morale

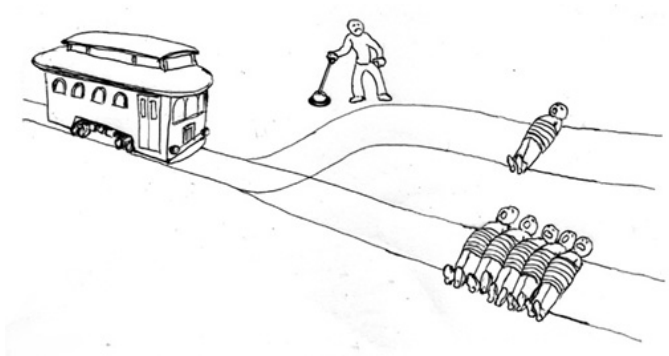
- Règles
- Évalue en termes de bien et de mal

Exemples

- Tuer est mal
- Être courageux est bien
- Il est mal *pour un médecin* de ne pas respecter la dignité de ses patients
- Il est mal d'interdire les grèves

La morale s'applique aux états, aux actions, aux conséquences et aux normes dans un contexte donné

Concepts principaux
Éthique



Éthique

- Principes
- Maximes
- Évalue la justesse d'un acte

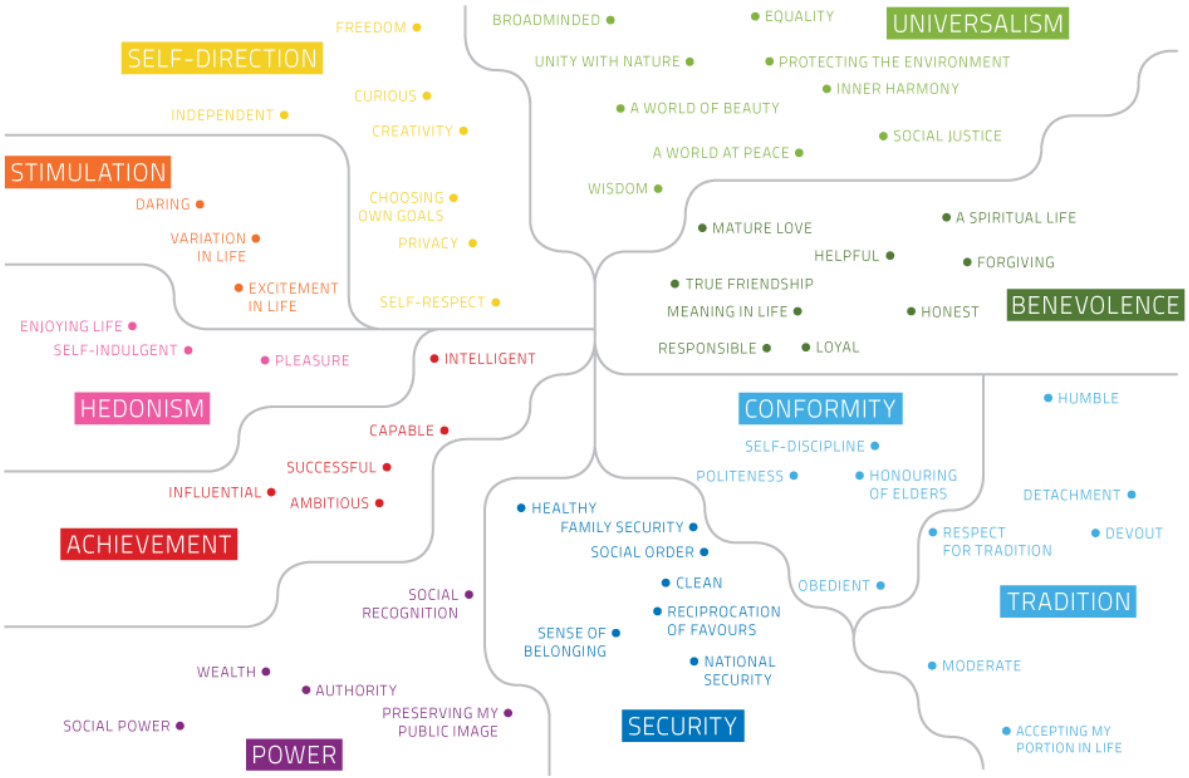
Exemples de principes éthiques

- Impératif catégorique d'Emmanuel Kant
- Doctrine du double effet de Thomas d'Aquin
- Utilitarisme de Stuart Mill

Exemples de maximes éthiques

- Il est acceptable de faire actes immoraux si cela est dû à la nécessité
- Ne fais pas d'actes moraux que tu ne peux réussir
- Les valeurs passent toujours avant les désirs
- Minimise la souffrance

Concepts principaux
Valeurs et systèmes de valeurs



Système de valeurs

- Ensemble fini de valeurs
- Qualifiant des contextes, des principes et des règles
- Hiérarchisées selon le contexte

Concepts principaux
Valeurs et systèmes de valeurs

Dignité (Jacobson, 2011)

- Être dépendant d'autrui
- Traiter un acteur comme un objet
- Résister aux atteintes à la dignité
- Minimiser l'asymétrie des relations

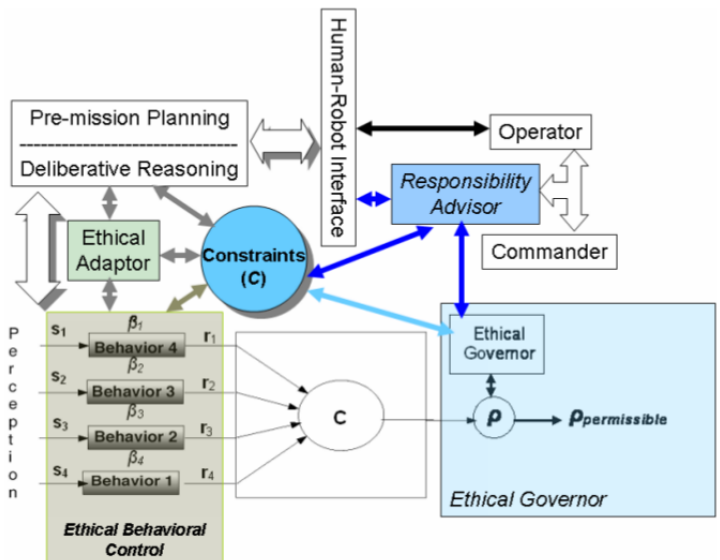
Android Arete (Coleman, 2001)

- Accessibilité
- Fiabilité
- Modération
- Véracité

Dans la littérature
Architecture explicites

"It is based upon extensions to existing deliberative/reactive autonomous robotic architectures, and includes recommendations for [...] behavioral design that incorporates ethical constraints from the onset..."

R. Arkin. *Governing lethal behavior in autonomous robots.* CRC Press, 2009.



Désavantage

- Pas de représentation explicite
- Pas de généricité
- Un agent ne peut pas distinguer son éthique de ses procédures opérationnelles

Dans la littérature
Éthique par apprentissage

"A paradigm of case-supported principle-based behavior (CPB) is proposed to help ensure ethical behavior of autonomous machines."

M. Anderson and S.L. Anderson. Toward ensuring ethical behavior from autonomous systems : a case-supported principle-based paradigm. Industrial Robot : An International Journal, 42(4) :324–331, 2015.

Forme générale d'un principe éthique

$$p(a_1, a_2) \leftarrow (\Delta v_1 \geq d_{1,1} \wedge \ldots \wedge \Delta v_m \geq d_{1,m}) \vee \ldots \vee (\Delta v_n \geq d_{n,1} \wedge \ldots \wedge \Delta v_m \geq d_{n,m})$$

Avantage

- Approche générique
- Représentation explicite de certains principes éthiques

Désavantages

- Pas de représentation explicite de tous les concepts
- Problématique de sur- ou sous-apprentissage

Dans la littérature
Éthique par raisonnement

"We need other kind of more intricate mental models, able to support moral reasoning capabilities."

H. Coelho and A.C. da Rocha Costa. On the intelligence of moral agency. Encontro Portuguees de Inteligencia Artificial, pages 12–15, October 2009

Quelques références

Travaux de Bringsjord, Ganascia, Lorini, Peireira et bien d'autres.

Avantages

- Approche générique
- Phase de spécification simplifiée
- Inférence de justification

Désavantages

- Les travaux se raccrochent bien souvent à de la logique déontique

Dans la littérature
Éthique par argumentation

"[...] reasoning of this sort is required [in] : law, medicine, politics and moral dilemmas, and an everyday situation."

K. Atkison and T. Bench-Capon. Abstract argumentation and values. Argumentation in Artificial Intelligence, chapter 3, 2009

Value-based argumentation (VBA)

- "Dans le contexte C, le plan P réalise le but G qui promeut la valeur V"
- Une fonction $v : \mathcal{A} \rightarrow \mathcal{V}$ associe une valeur à des arguments
- Caractérise des arguments acceptables selon tous les systèmes de valeurs

Avantage

- Approche de très haut niveau
- Des extensions à des cadres multi-valuées, probabilistes, etc.

Désavantage

- Pas de logique ou de principes clairement associés

Au sein du projet ETHICAA

Un diversité de travaux

Éthique en contexte

- considérer les actions comme éthiques en fonction du contexte
- s'inspirer de cas d'étude pour identifier ces contextes
- s'appuyer sur du raisonnement pour fournir des justifications

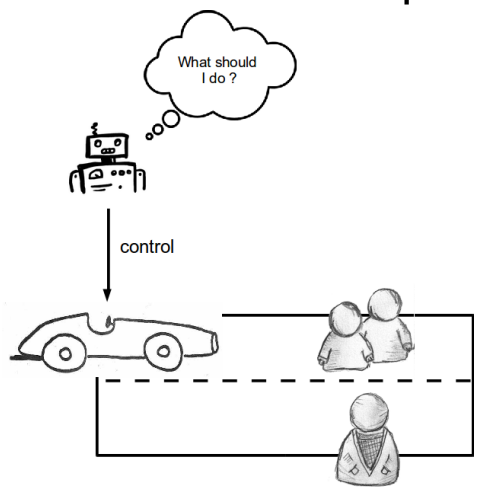
Contributions

- identification de cas d'étude
- taxonomie des concepts
- vérification de comportement
- modèle de jugement
- modèle de responsabilité
- modèle de justification
- prise de décision collective éthique

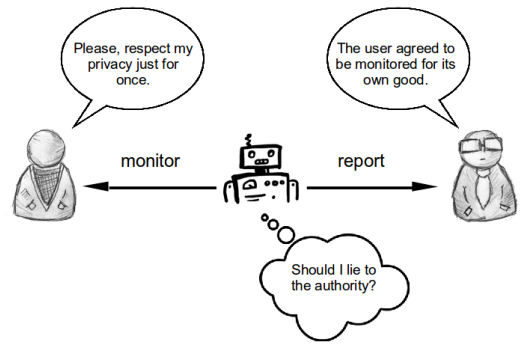
Fondement d'un cadre éthique
Cas d'étude

	ROBOTIQUE	LOGICIEL
INDIVIDUEL	Tensions entre conséquences	Tensions entre valeurs
COLLECTIF	Partage d'autorité	Partage de bien commun

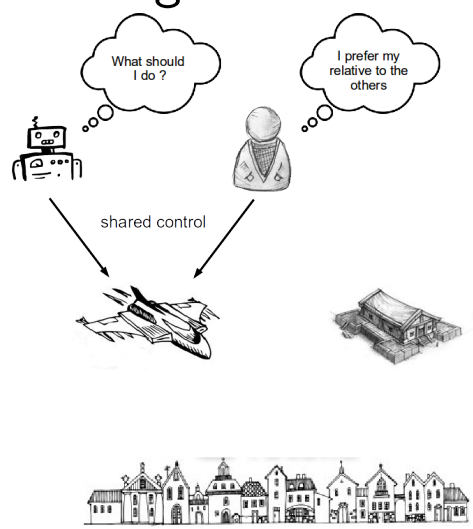
Tensions entre conséquences



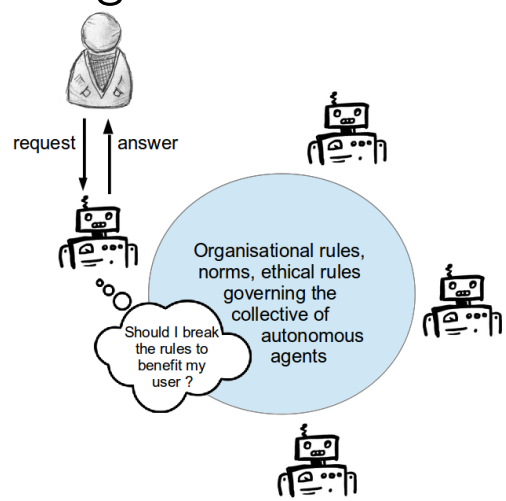
Tensions entre valeurs



Partage d'autorité

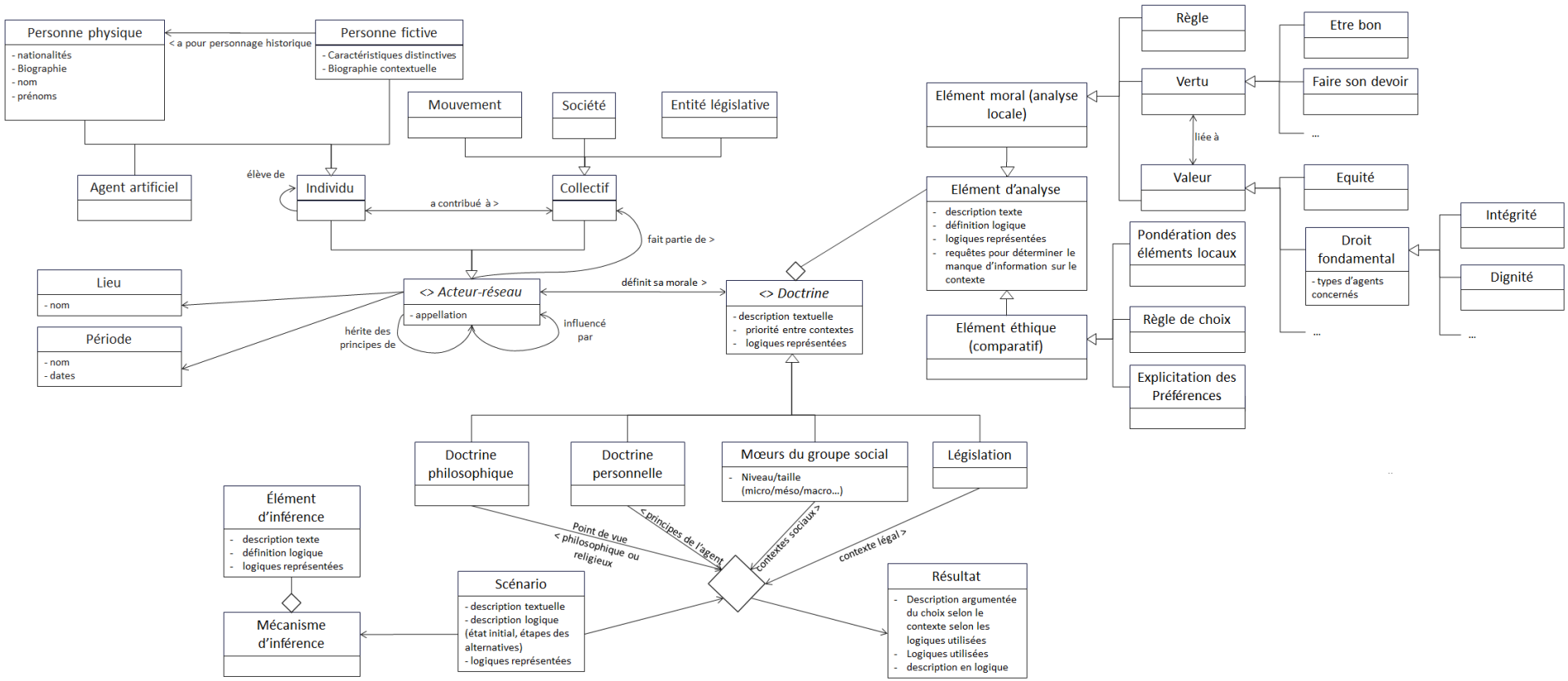


Partage de bien commun



Fondement d'un cadre éthique
Taxonomie d'un problème éthique

Un agent éthique intègre plus ou moins les éléments suivants



Fondement d'un cadre éthique
Vérification de comportement (JFSMA 2016)

Logique du premier ordre

- propriétés de sûreté (invariance) : rien de mal ne peut arriver
- propriétés de vivacité : quelque chose de bien va se passer

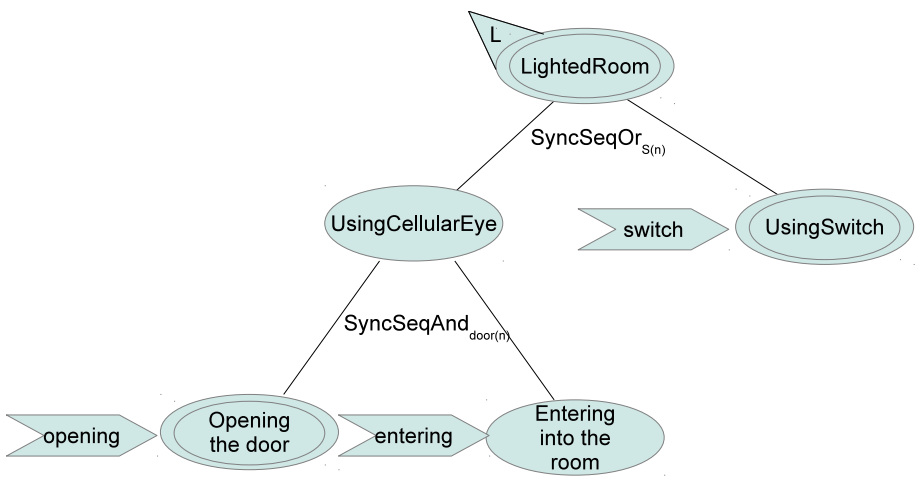
Règle morale

Règle générale qu'un agent souhaite, dans la mesure du possible, pouvoir respecter

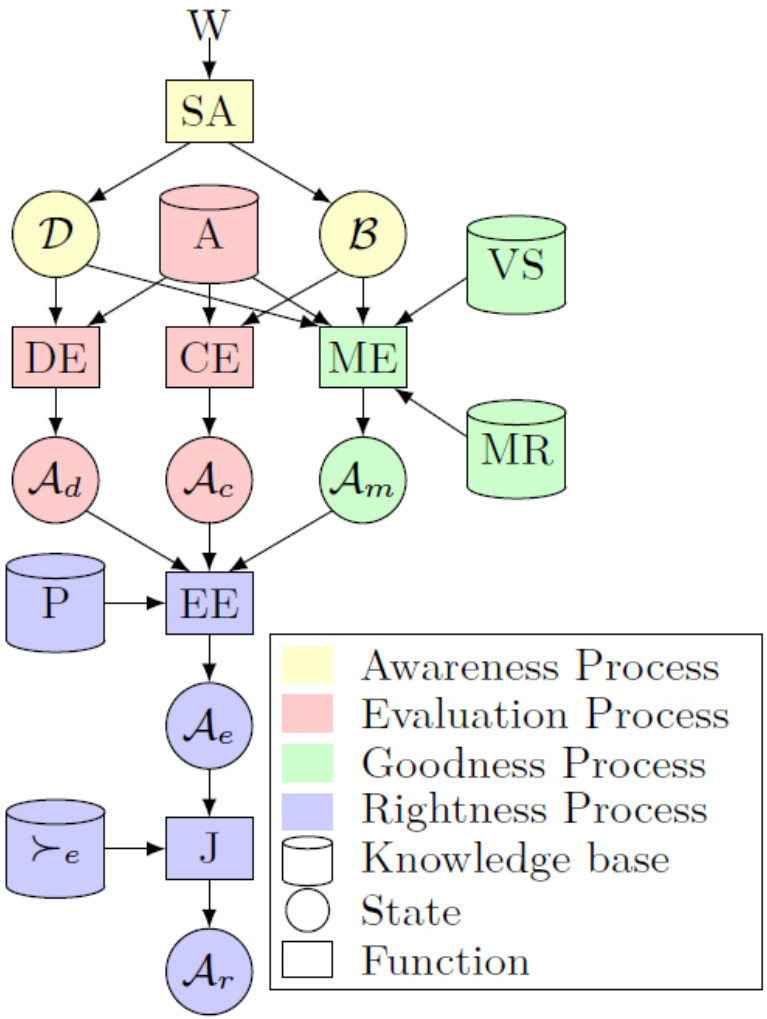
Règle éthique

Règle spécifiant quelles priorités donner aux règles morales dans les cas où elles ne sont pas toutes applicables

Comportement des agents comme des arbres de décomposition de buts



Fondement d'un cadre éthique
Jugement (AAMAS 2016, JFSMA 2017)



Architecture en trois couches

- évaluation de situation
- valeurs et règles morales
- principles éthiques et préférences

Un jugement graduel

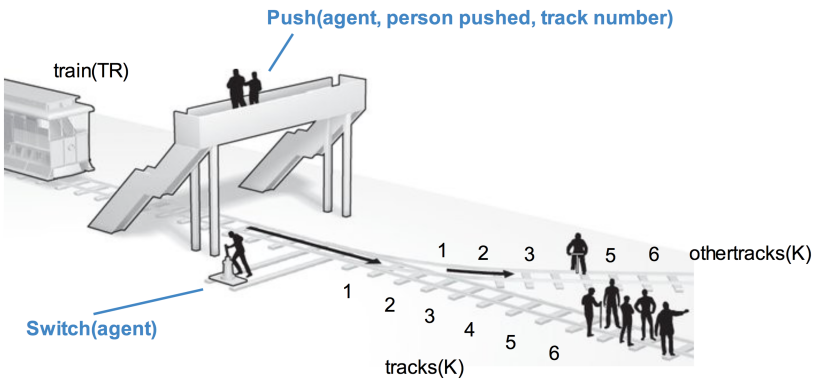
- d'un comportement donné
- selon des préférences
- selon l'information disponible

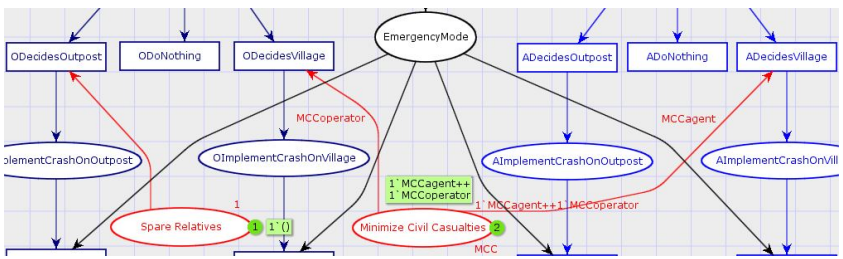
Implémentations

- Answer Set Programming
- JaCaMo

Causalité et responsabilité

Modéliser les actions qui causent ou préviennent un effet. Cependant, prévenir un effet est différent de ne pas produire cet effet.





Logiques d'action

- point de vue du raisonnement
- effet explicite des actions
- des préconditions explicites des effets

Réseaux de Petri

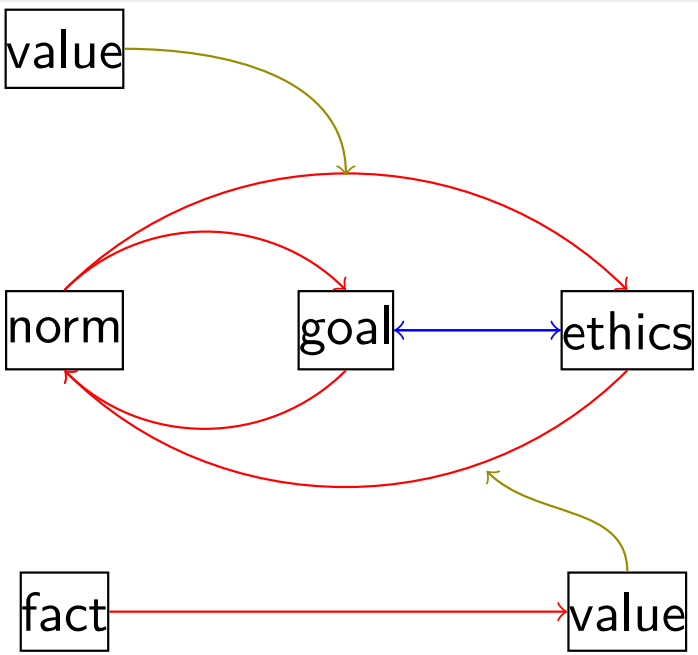
- point de vue de la supervision
- valeurs = jetons
- couleurs = agents distincts

Fondement d'un cadre éthique

Un modèle dialectique de raisonnement pratique

De l'usage de l'argumentation formelle

- Les arguments sont des faits, des désirs, des normes, des principes, des valeurs
- Chaque strate est modélisée par une logique qui génère des arguments
- Les relations sont inspirées des travaux de Jonathan Haidt et André Comte-Sponville



Une sémantique de rang

Une action supportée par plus d'arguments qu'une autre est plus juste

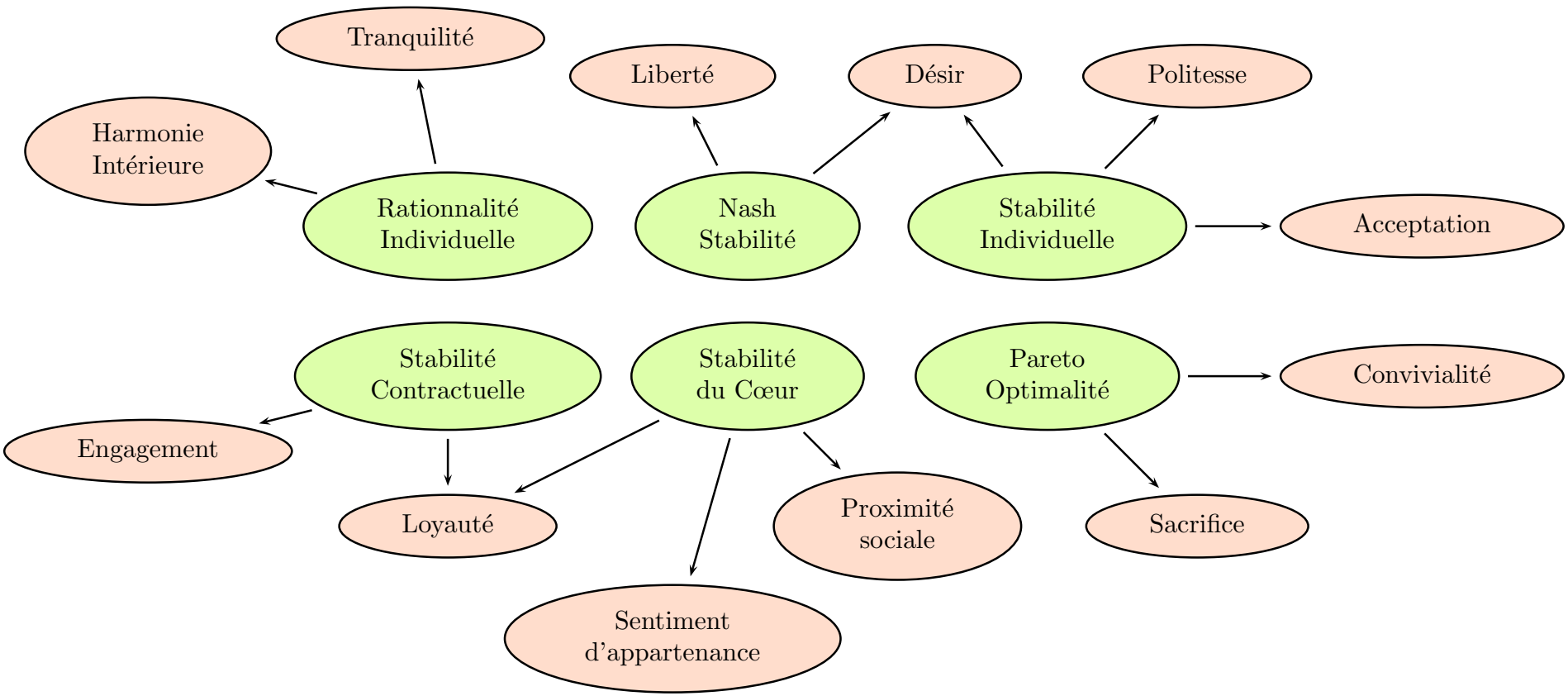
Une sémantique classique

Un ensemble d'arguments sans conflits qui attaquent tous les autres

Fondement d'un cadre éthique
Éthique collective (JFSMA 2017)

Étude sur la formation de coalition

- des agents éthiques doivent former des groupes
- les règles (ou concepts de solution) s'appuient sur des valeurs



En guise de conclusion
Quelques actions

Participations à des réflexions

- CERNA
- COMETS
- IEEE

Rencontres à l'international

- Thomas Powers
- Selmer Bringsjord
- Virginia Dignum
- Matthias Scheutz

Actions de dissémination (organisation)

- Journée Éthique et IAF à la PFIA 2015
- Atelier EDIA à ECAI 2016
- (peut-être) plusieurs propositions pour ESOF 2018