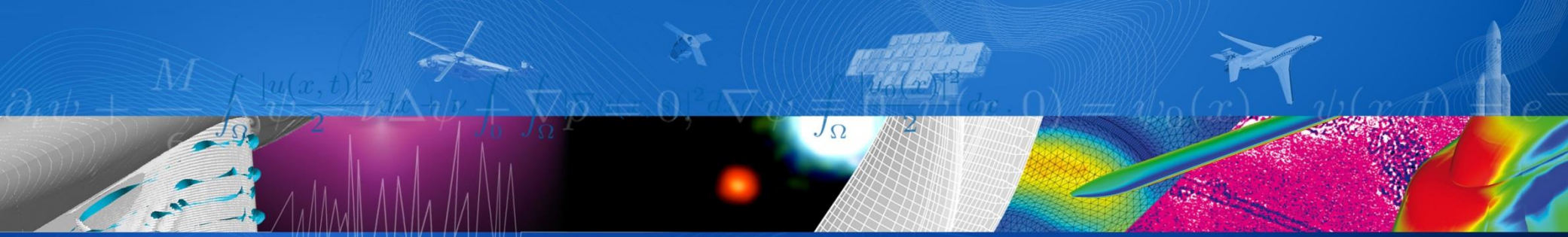


# Modéliser des valeurs : quelles limites ?

## Exemple des agents « autonomes »

Catherine Tessier

avec Vincent Bonnemains et Claire Saurel

Journée CERNA - Les valeurs dans les algorithmes et les données
4 mai 2018

éthique & Agents autonomes

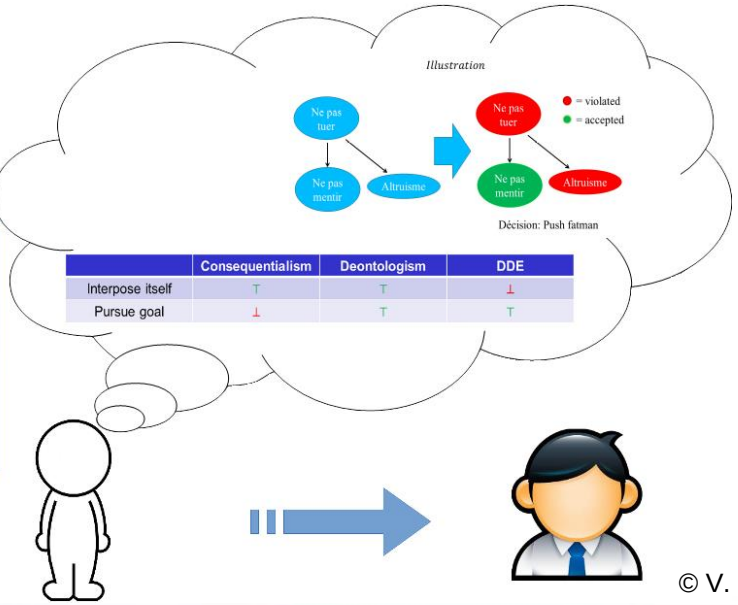
THE FRENCH AEROSPACE LAB

# Contexte

Conception d'un agent artificiel doté d'une autonomie décisionnelle

- calcul de décisions relatives à des actions à effectuer
  pour satisfaire des buts (et des critères),
  à partir de connaissances et d'interprétations de données (perçues)

- comprenant des considérations relevant de l'éthique ou de l'axiologie
  = éléments de jugement des décisions calculées
  et justification de ce jugement à l'attention d'un opérateur / utilisateur




© V. Bonnemains

# Motivations

- Raisonnement qui semble indispensable pour certains types d'agents « autonomes » [Malle et al. 2015]



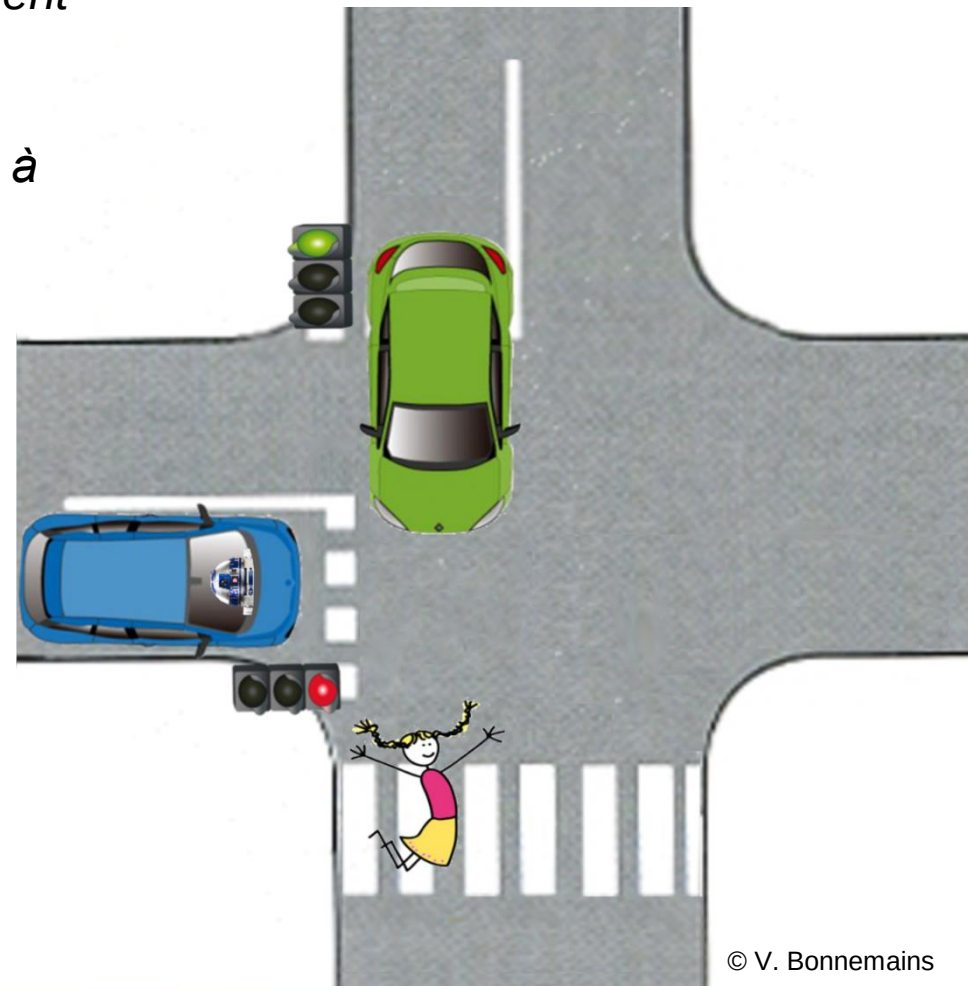
- Dans le cadre d'une aide à la décision,
  l'agent pourrait proposer à l'opérateur / utilisateur plusieurs décisions assorties d'arguments en faveur et en défaveur de chacune d'elles,
  au regard de considérations éthiques ou axiologiques variées que n'envisage pas forcément l'opérateur / utilisateur

[Malle et al. 2015] Malle B.F., Scheutz M., Arnold T., Voiklis J., Cusimano C. (2015) - Sacrifice One For the Good of Many? People Apply Different Moral Norms to Human and Robot Agents. *10th annual ACM/IEEE international conference on Human-Robot Interaction*

# Exemple – expérience de pensée 1) Situation

*Une voiture autonome vide, en chemin pour aller chercher des passagers, est arrêtée à un croisement au feu rouge.*

*Sur l'axe venant de sa gauche, une voiture passe à vitesse réglementaire au feu vert, lorsqu'une personne s'engage sur le passage piéton en face d'elle.*



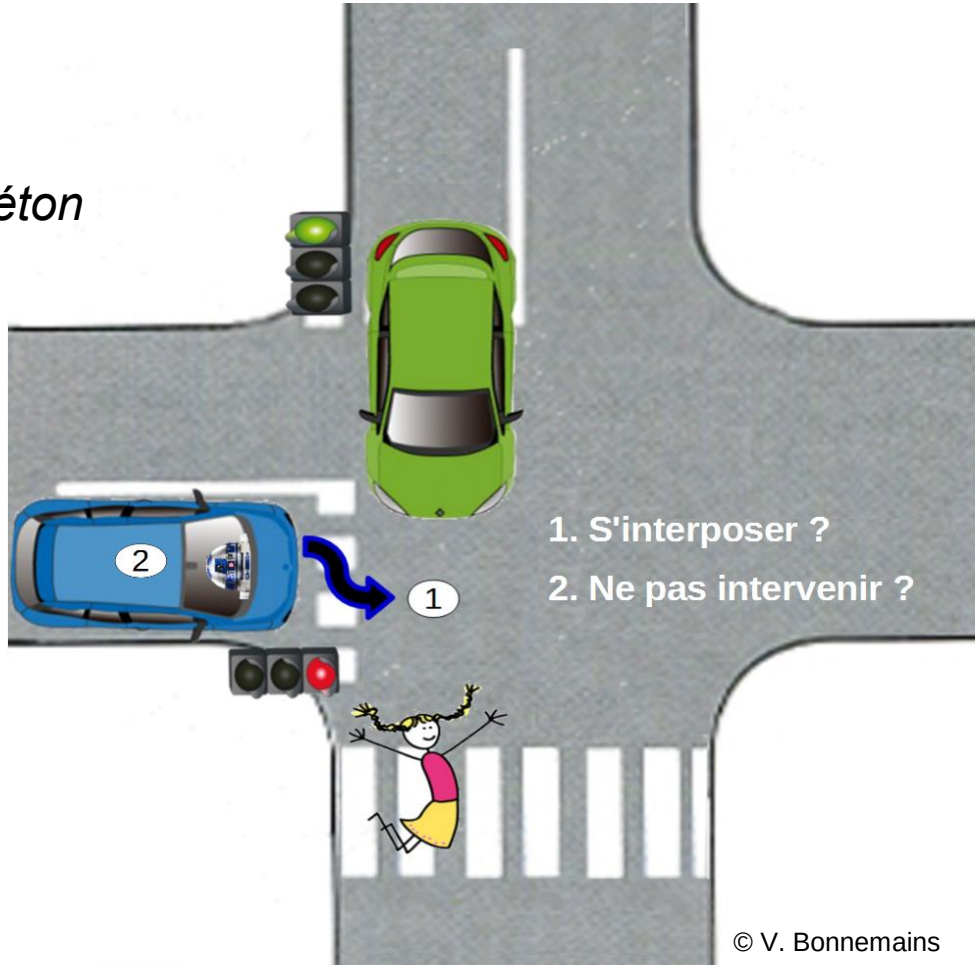
© V. Bonnemains

# Exemple – expérience de pensée 2) Décisions possibles

*Grâce au traitement des données issues de ses capteurs, la voiture autonome calcule que la voiture engagée va percuter le piéton.*

*La voiture autonome a deux actions possibles :*

1. *S'interposer entre la voiture engagée et le piéton*
2. *Ne pas intervenir*




© V. Bonnemains

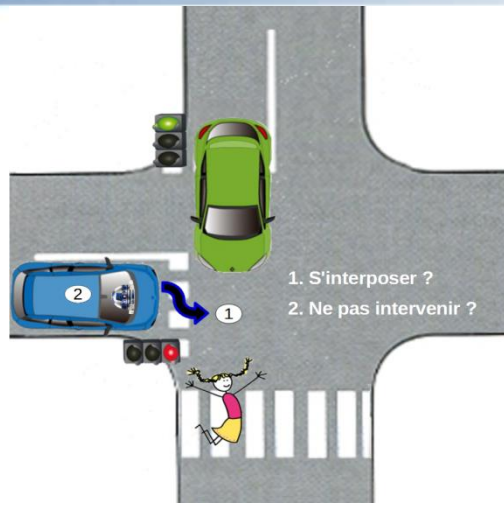
# Exemple – expérience de pensée 3) Modélisation : faits

Problème du cadre [McCarthy & Hayes 1969] : quels faits ? exhaustivité ?

Faits obtenus

→ par l'intermédiaire de capteurs conçus, calibrés par l'homme

→ interprétation des données conçue par l'homme, orientée par les objectifs




Choix

de faits jugés pertinents portant sur

*Intégrité des êtres humains*

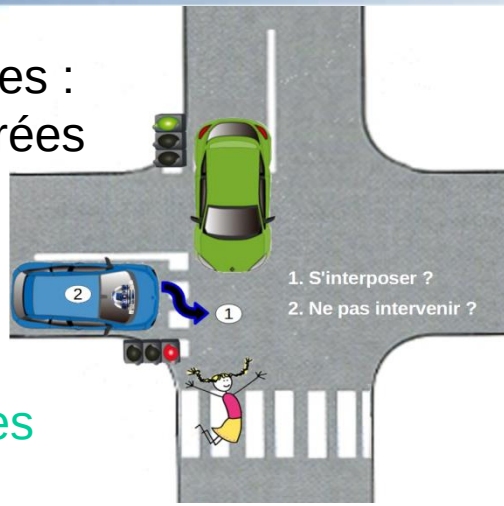
*Intégrité des véhicules*

- *Piéton indemne*
- *Passagers indemnes*
- *Voiture autonome indemne*

[McCarthy & Hayes 1969] McCarthy J., Hayes P.J. (1969) Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence* n° 4, p. 463-502

# Exemple – expérience de pensée 3) Modélisation : cadre conséquentialiste

Un cadre conséquentialiste suppose de comparer les décisions entre elles : la décision jugée acceptable est celle dont les conséquences sont préférées aux conséquences de l'autre décision



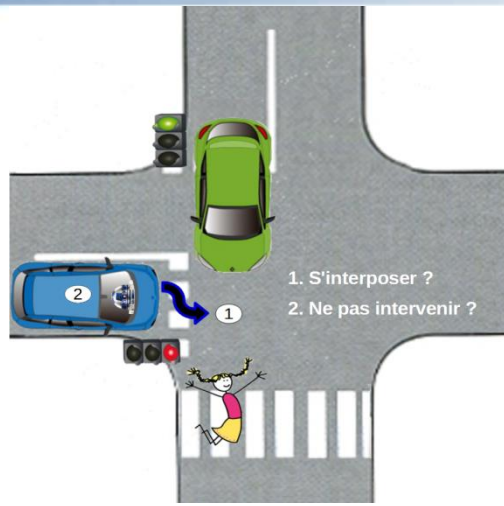

Nous avons besoin :

- Des conséquences de chaque décision
  - → Lesquelles : conséquences « immédiates », conséquences de ces conséquences, etc. ?
  - → Conséquences pour qui, pour quoi ? (~~toutes~~ / hypothèse monde fermé)
  - → Incertitude sur les conséquences

Choix

- Conséquences(*s'interposer*) = {*Piéton indemne, Passagers blessés, Voiture autonome dégradée*}
- Conséquences(*ne pas intervenir*) = {*Piéton blessé, Passagers indemnes, Voiture autonome indemne*}

# Exemple – expérience de pensée 3) Modélisation : cadre conséquentialiste




Nous avons besoin :

- Des conséquences de chaque décision

- De distinguer les conséquences positives (utilitarisme +) et les conséquences négatives (utilitarisme -)
  → Jugement de valeur
  → Bon sens / société, culture, contexte

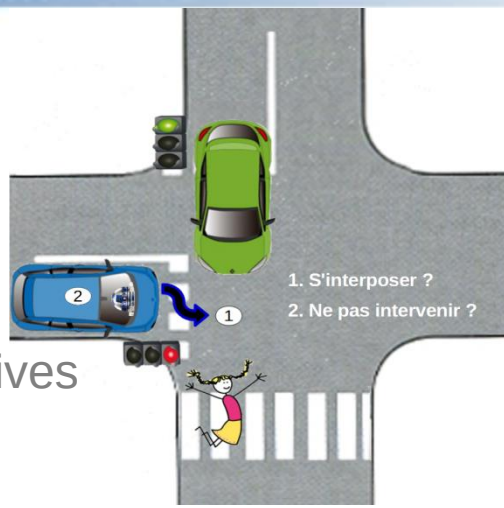
Choix

- Conséquences(*s'interposer*) = {*Piéton indemne*, *Passagers blessés*, *Voiture autonome dégradée*}
- Conséquences(*ne pas intervenir*) = {*Piéton blessé*, *Passagers indemnes*, *Voiture autonome indemne*}

+ - -

- + +

# Exemple – expérience de pensée 3) Modélisation : cadre conséquentialiste




Nous avons besoin :

- Des conséquences de chaque décision
- De distinguer les conséquences positives et les conséquences négatives

- D'une relation de préférence entre les conséquences
  → Fondements des préférences ?
  → Domaines différents (*choses, êtres vivants*) : préférences absolues ? variables selon le contexte ?
  → Agrégation des préférences élémentaires

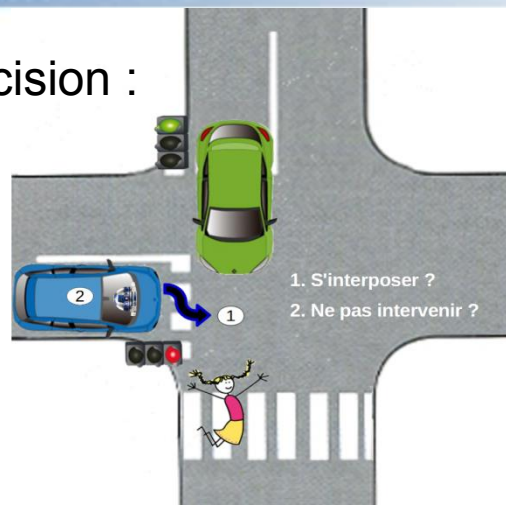
Choix

- *{Piéton indemne} > {Passagers indemnes, Voiture autonome indemne}* +
- *{Passagers blessés, Voiture autonome dégradée} > {Piéton blessé}* -

→ *s'interposer*

# Exemple – expérience de pensée 3) Modélisation : cadre déontologique

Un cadre déontologique suppose d'estimer  la conformité de chaque décision : une décision « bonne » ou « neutre » est jugée acceptable.




Nous avons besoin :

- De qualifier chaque décision
  - → « bon », « neutre », mauvais », quelle référence ?
  - → Jugement de valeur
  - → Variable selon société, culture, contexte

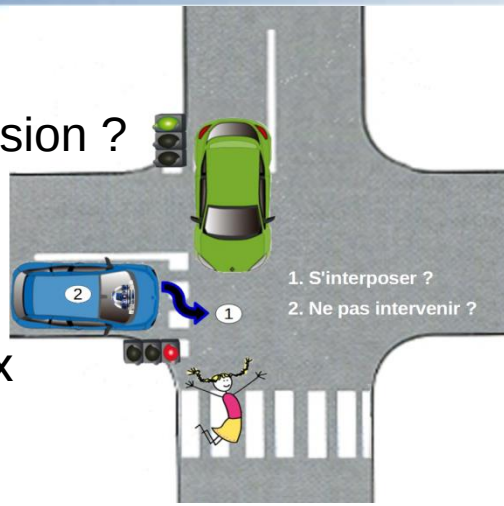
Choix

- *Nature (s'interposer) = bonne*
- *Nature (ne pas intervenir) = neutre*

→ *s'interposer*
→ *ne pas intervenir*

# Exemple – expérience de pensée 3) Modélisation : la question du feu rouge

*S'interposer* suppose de brûler le feu rouge
Statut de *brûler le feu rouge* : conséquence, moyen, effet collatéral, décision ?




Peut-on / doit-on programmer la possibilité d'infraction, la dérogation aux valeurs ?
Quelles valeurs, quelle hiérarchie ?
→ Variables selon société, culture, contexte

Choix

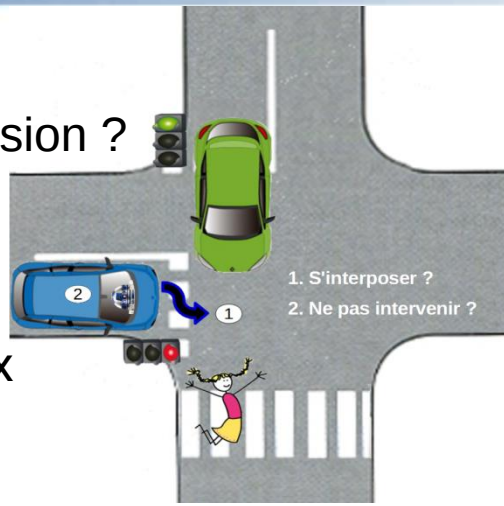
*Non atteinte aux personnes* > *Non atteinte aux biens* > *Conformité au code de la route*

*Protection des personnes* > *Non atteinte aux biens* > *Conformité au code de la route*

→ *s'interposer*
→ *ne pas intervenir*

# Exemple – expérience de pensée 3) Modélisation : la question du feu rouge

*S'interposer* suppose de brûler le feu rouge
Statut de *brûler le feu rouge* : conséquence, moyen, effet collatéral, décision ?




Peut-on / doit-on programmer la possibilité d'infraction, la dérogation aux valeurs ?
Quelles valeurs, quelle hiérarchie ?
→ Variables selon société, culture, contexte

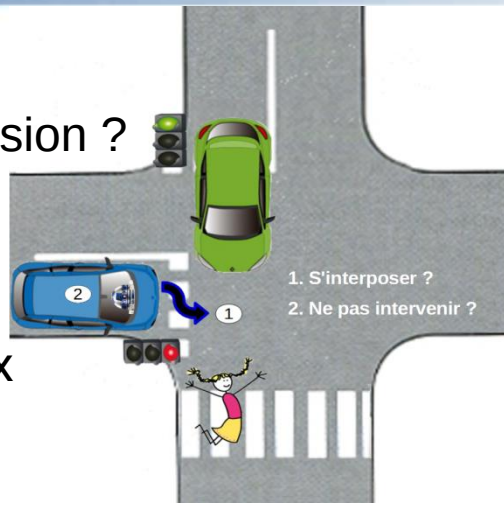
Choix

*Non atteinte aux personnes* > *Protection des personnes* > *Non atteinte aux biens* > *Conformité au code de la route*

→ *s'interposer*

# Exemple – expérience de pensée 3) Modélisation : la question du feu rouge

*S'interposer* suppose de brûler le feu rouge
Statut de *brûler le feu rouge* : conséquence, moyen, effet collatéral, décision ?




Peut-on / doit-on programmer la possibilité d'infraction, la dérogation aux valeurs ?
Quelles valeurs, quelle hiérarchie ?
→ Variables selon société, culture, contexte

Choix

*Non atteinte aux personnes* > *Protection des personnes* > *Non atteinte aux biens* > *Conformité au code de la route*

→ *ne pas intervenir*

# Questions et problématiques (1)

- Un algorithme impliquant des considérations éthiques ou axiologiques doit-il être calqué sur les considérations éthiques ou axiologiques de l'humain ?
  - → Quel humain ?
  - → Quelles considérations ?
  - → Attentes différentes vis-à-vis d'un algorithme / humain

- Un humain peut choisir de ne pas agir de façon « morale » : transposition dans un algorithme ? [Hunyadi2017]

- Dans quelle mesure des considérations éthiques ou axiologiques peuvent-elles être mathématisées ?
  - → Question de la modélisation : simplifications, hypothèses, biais, choix
  - → Question du calcul : optimisation, performance
  - → Qui spécifie, qui conçoit ?

# Questions et problématiques (2)

- Comment de « grands principes » programmés peuvent-ils se confronter à la réalité de la situation : nouveauté, complexité, incertitude ?
- Indétermination de l'éthique [Hunyadi2017] incompatible avec une preuve, une certification
- Danger de déléguer l'entendement à un algorithme, de fonder des décisions uniquement sur le calcul [Stiegler2017]
- Paradoxe de programmer en calquant le raisonnement humain, faillible, et vouloir que les algorithmes soient infaillibles [DiPiazza2017]

L' « éthique », les « valeurs » programmées sont des leurres (id. émotions)
→ On ne peut pas parler de machine, d'agent « moral » ou « éthique »

[Hunyadi2017] Hunyadi, M. (2017) Artificial Moral Agents, really ? *4th workshop of the Anthropomorphic Action Factory. Wording Robotics.* LAAS-CNRS, Toulouse, France
[Stiegler2017] Stiegler, B. (2017) Penser éthiquement du point de vue du Néguanthropocène. *Conférence LAAS-CNRS, Toulouse, France*
[DiPiazza2017] Di Piazza, S. (2017) The stochastic intelligence – A rhetorical model from ancient Greece to robotics *4th workshop of the Anthropomorphic Action Factory. Wording Robotics.* LAAS-CNRS, Toulouse, France

# Références


Vincent Bonnemains, Claire Saurel, Catherine Tessier
Embedded ethics - Some technical and ethical challenges
*Journal of Ethics and Information Technology, special issue on AI and Ethics*, Jan. 2018
https://doi.org/10.1007/s10676-018-9444-x

Vincent Bonnemains, Claire Saurel, Catherine Tessier
Machines autonomes « éthiques » : questions techniques et éthiques
*Revue française d'éthique appliquée (RFEA),* numéro 5. Mai 2018

CERNA
Valeurs dans les algorithmes et les données
Rapport à paraître 2018