

Questionnements éthiques sur la robotique et l'intelligence artificielle

Raja Chatila

ISIR, Université Pierre-et-Marie-Curie

Journée « Philosophie des sciences et intelligence artificielle »
le 2 février 2017, à l'École normale supérieure



Organisée par l'AFIA, la SPS et le DEC

Responsables scientifiques :

Robin Lamarche-Perrin et Daniel Andler



AfIA
Association française
pour l'intelligence Artificielle



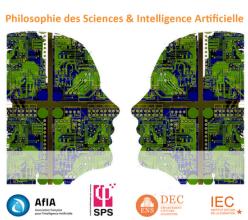
SPS
SOCIÉTÉ DE PHILOSOPHIE DES SCIENCES



DEC
DÉPARTEMENT
D'ÉTUDES
COGNITIVES



IEC
INSTITUT
DE LA COGNITION ★



Quelques Questionnements Ethiques sur la Robotique et l'Intelligence Artificielle

Raja Chatila

Institut des Systèmes Intelligents et de Robotique (ISIR)
CNRS et Université Pierre et Marie Curie, Paris

Membre de la CERNA, *Commission de réflexion sur l'éthique de la recherche en sciences et technologies du numérique d'Allistene*

Chair, *The IEEE Global Initiative on Ethical Considerations in AI and Autonomous Systems*

Le robot



- Un robot est une **machine matérielle** munie des capteurs et d'actionneurs et contrôlée par des programmes (des algorithmes) exécutés par des ordinateurs.
- Elle possède des capacités de **perception** de son environnement, d'**action** et de mouvement, ainsi que de prise de **décision** pour effectuer des tâches.
- Elle peut aussi posséder des capacité de communication et **d'interaction**.
- Elle peut aussi **apprendre** des actions, des représentations.
- Ces capacités peuvent être développées à divers niveaux de complexité et confèrent aux robots des degrés d'autonomie différents.

Le robot comme **intelligence artificielle « encorporée » confrontée au réel**.

Constat

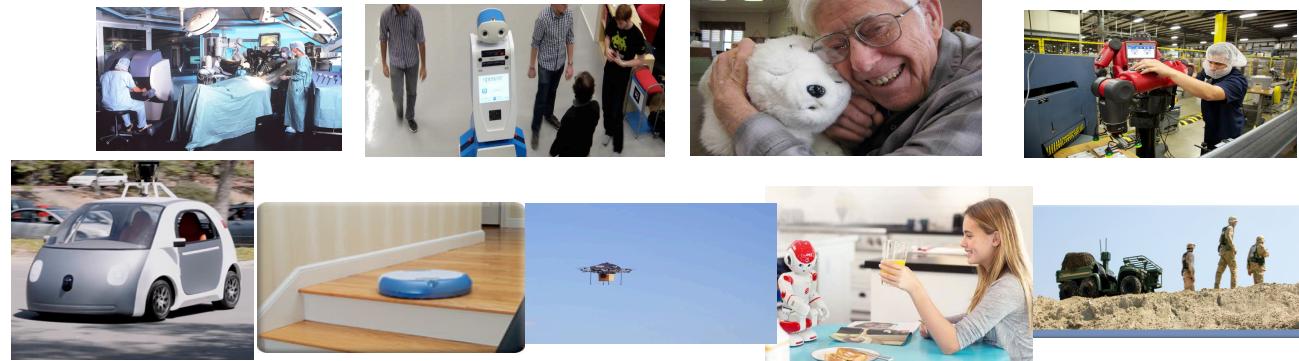
- La robotique et l’Intelligence Artificielle ont progressé de manière accélérée ces cinq dernières années.
 - Nouvelles méthodes pour la perception, la reconnaissance, la décision l’apprentissage, liées à la disponibilité de grandes quantités de données.
 - Favorisées par la vitesse des processeurs, la disponibilité de la mémoire et des données, la miniaturisation des capteurs, des actionneurs et des sources d’énergie.
- Des applications deviennent possibles dans plusieurs domaines.

Applications: production industrielle, milieux dangereux, transport, agriculture, construction, santé, loisirs, défense,...

- Remplacer les humains



- Assister et servir les humains



- Rehabiliter/augmenter les humains



Prise de conscience éthique en R&IA

- La *recherche responsable* est un mouvement qui s'est développé en Europe depuis plusieurs dizaines d'années.
- Les questionnements sur les problèmes éthiques, légaux et sociaux (ELS) spécifiquement dans l'usage des robots et le développement de l'IA datent d'une quinzaine d'années.
- Préoccupations récentes concrètes sur les dangers potentiels de l'IA et la robotique (décisions algorithmiques, emploi, armes autonomes...).

Remarque:

L'état de l'art réel est souvent en deçà des questions posées.

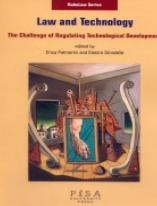
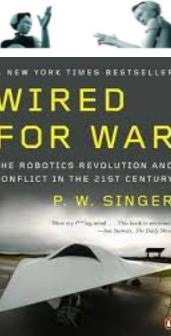
Quelques sujets "Ethiques, Légaux, Sociétaux" (ELS)

- Economie, robotisation, emploi
- Protection de la vie privée
- Surveillance
- Dignité humaine
- Autonomie humaine
- Dépendance, isolation
- Liens affectifs avec les robots
- Augmentation humaine
- Intégrité et identité humaine
- Imitation du vivant
- Ethique de l'usage des robots
- Prise de décision autonome
- Responsabilités morale et juridique
- Statut du robot dans la société



**ROBOT
ETHICS**

Edited by
Patrick Lin, Kristin阿明,
and George A. Bekey



Ethique du chercheur/concepteur, de l'usage, de la machine

- Trois facettes liées:
 - Ethique de la *recherche*: recherche responsable. Respect de préconisations ou de règles par le concepteur, le chercheur; Méthodologie de conception intégrant des valeurs.
 - Ethique de l'*usage*: mise en œuvre respectant des règles éthiques.
 - Ethique des *systèmes ou des machines* : règles ou comportements éthiques inclus dans le fonctionnement du système (agents moraux artificiels).

Capacités posant des questions éthiques

- Autonomie, capacités décisionnelles, apprentissage
- Interactions affectives et sociales
- Imitation du vivant
- Réparation et augmentation de l'humain

Autonomie

Autonomie: Capacité d'un agent à prendre des décisions sans l'assistance d'un autre agent.

- Autonomie opérationnelle: Nécessaire dans tous les systèmes
 - Concerne le traitement des données et leur interprétation, les décisions simples, la commande et l'exécution des actions.
- Autonomie décisionnelle: Concerne une interprétation plus sémantique, l'évaluation de situations et la prise de décisions non triviales.

Continuum entre le système automatique et le système autonome

Exemples d'autonomie



Boston Dynamics

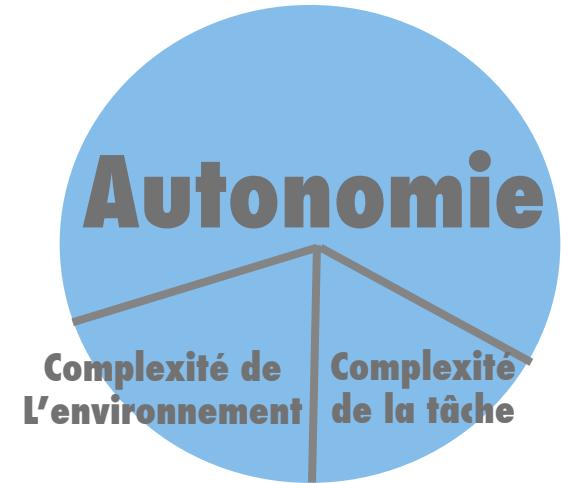
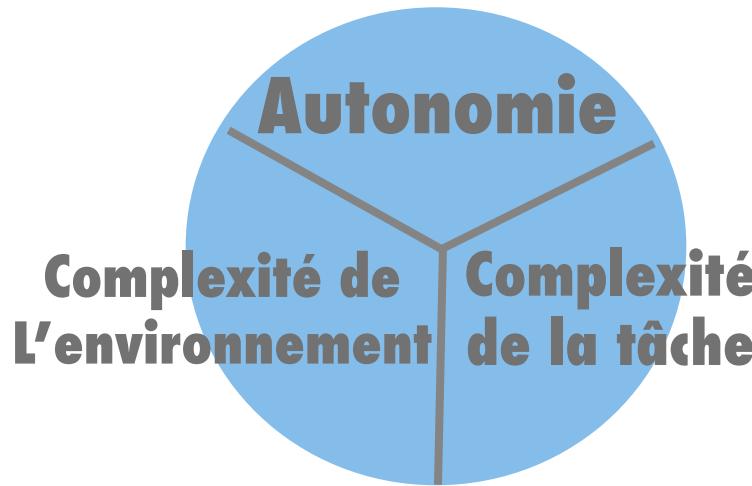
Contrôle automatique avancé



Autonomie opérationnelle.



L'autonomie est un concept relatif



- L'autonomie atteignable dépend de la complexité de l'environnement et de celle de la tâche.
- La complexité de l'environnement peut être mesurée par la quantité d'information et son flux.
- La complexité de la tâche dépend de la dimension et de la structure de l'espace d'état du processus de décision.

Systèmes d'armes automatiques/autonomes



Phalanx
(US D.o.D)



Samsung SGR-A1

Autonomie et partage d'autorité

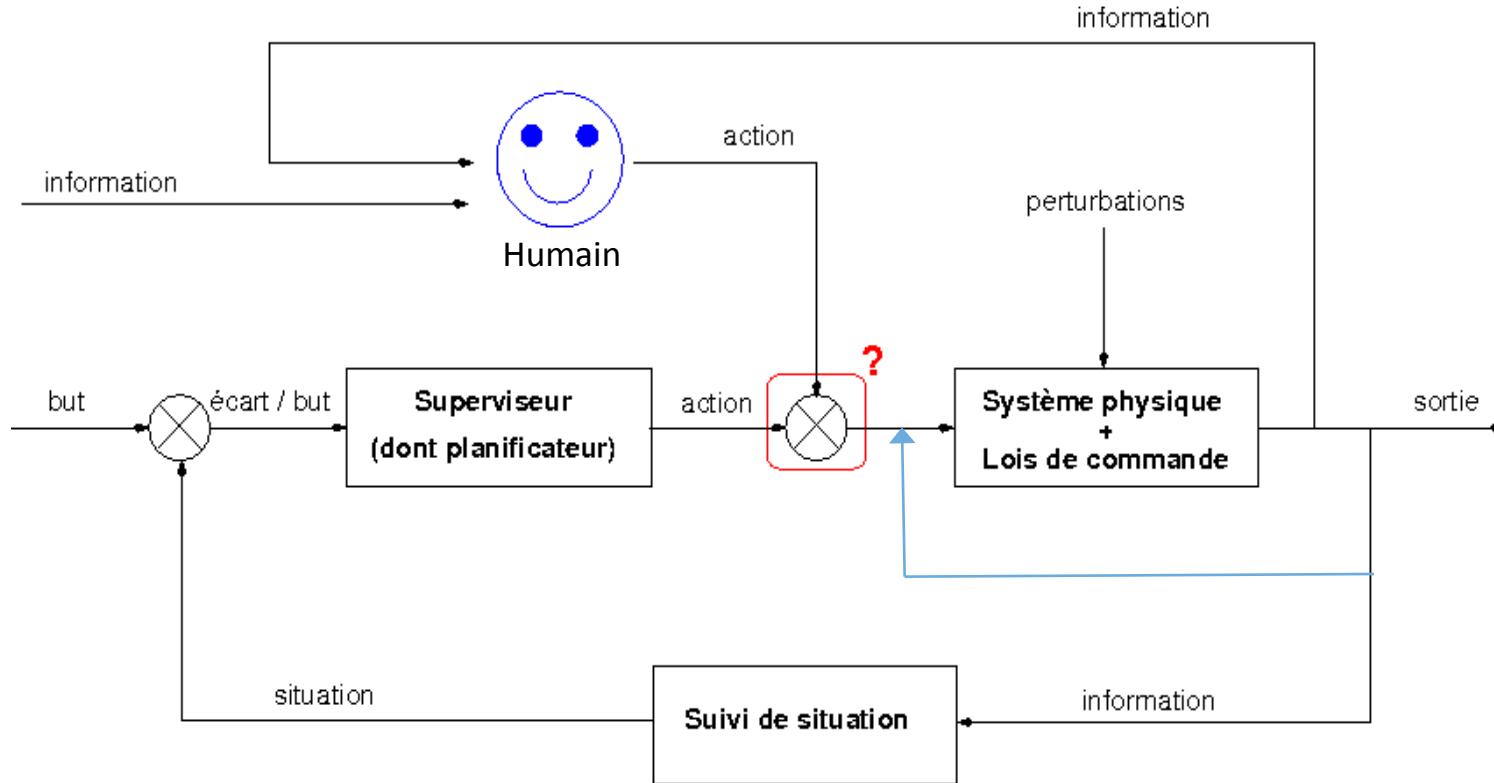
Système avec humain dans la boucle

How drones work



Drone Reaper

Partage d'autorité: L'opérateur et la machine peuvent décider des actions



Autonomie et partage d'autorité

- Machine
 - Capacités de décision limitées
 - Incertitudes de perception et d'action
 - Interprétation de situation limitée
 - Rapidité et réactivité
 - Efficacité
- Humain
 - Attention limitée
 - Perception limitée
 - Stress et émotions
 - Réalisation de scénario
 - Conscience de situation
 - Jugement moral

Interaction

Problèmes de communication

Biais d'automation: confiance excessive en la machine

Surprises: ignorance de l'état exact de la machine en cas de reprise

Buffer moral: responsabilité du robot/responsabilité de l'homme

Système Aegis sur le USS Vincennes

Incident de l'Iran Air 655 (3/07/1988)



Aegis Data Report:

- Iran Air Flight 655 continuously ascended in duration of flight
- Iran Air Flight 655 continuously squawked Mode III identification, friend or foe (IFF) in duration of flight
- Iran Air Flight 655 held consistent climb speed in duration of flight

Fogarty, William M. (July 28, 1988).

"Formal Investigation into the Circumstances

Surrounding the Downing of Iran Air Flight 655 on 3 July 1988 ».

93-FOI-0184. Archived from the original on 6 May 2006.

Retrieved March 31, 2006. (Wikipedia)

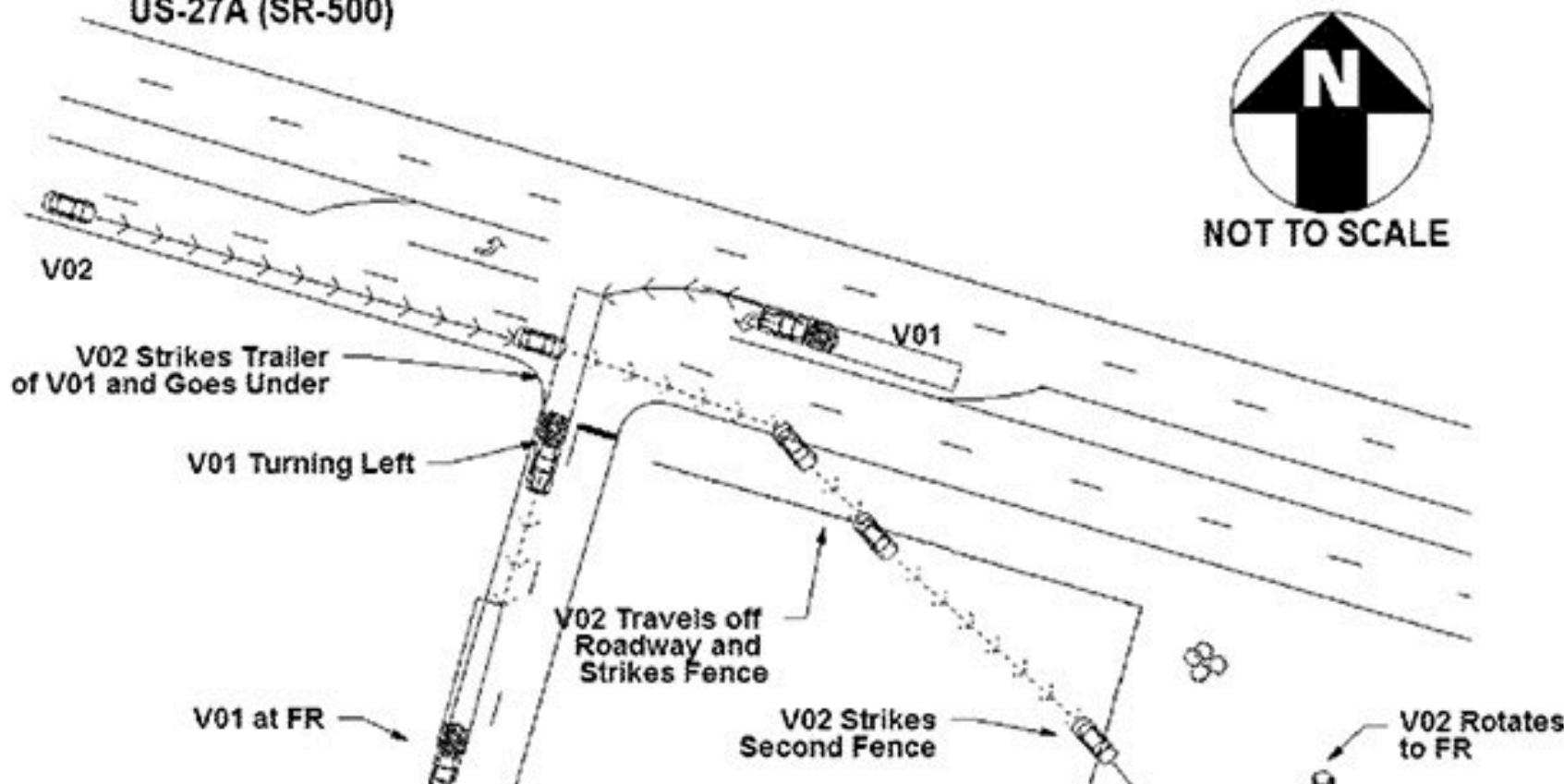
Personnel Report To Commanding Officer:

- Iran Air Flight 655, after attaining 9,000 to 12,000 ft (2,700 to 3,700 m), reportedly descended on an attack vector on USS Vincennes
- Iran Air Flight 655 reportedly squawked Iranian F-14 Tomcat on Mode II IFF for a moment; personnel proceeded to re-label the target from "Unknown Assumed Enemy" to "F-14 »
- Iran Air Flight 655 was reported to increase in speed to an attack vector similar to an F-14 Tomcat

Accident Tesla Autopilot

Date of Crash 07/May/2016 04:40 PM	Date of Report 07/May/2016 04:40 PM	Invest. Agency Report Number FHPB16OFF012208	HSMV Crash Report Number 85234095
---------------------------------------	--	---	--------------------------------------

US-27A (SR-500)



Ethique de la machine

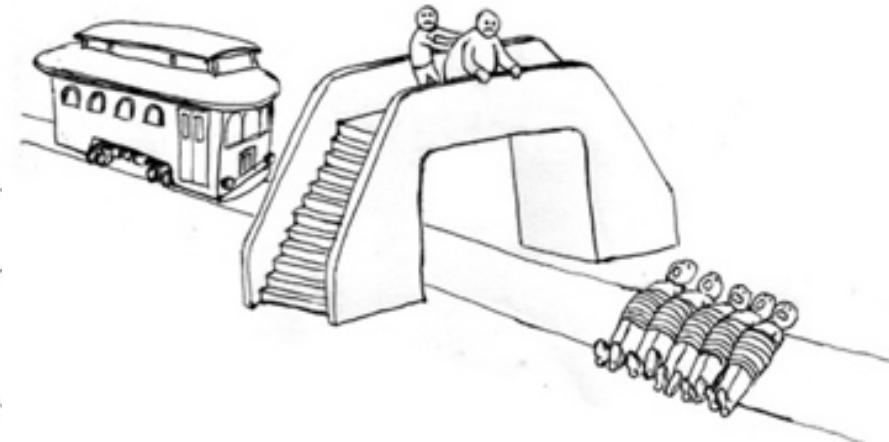
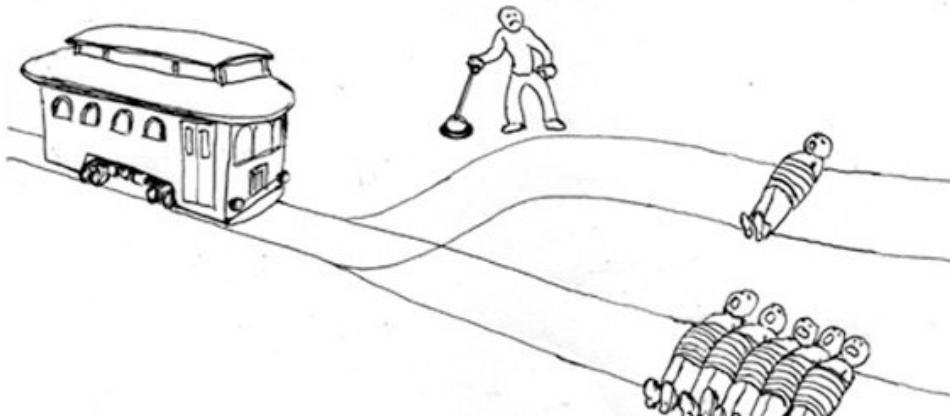
- Le système respecte des valeurs morales dans ses décisions
- Les théories éthiques adoptées déterminent les valeurs respectées
- Difficulté: Définition des valeurs; évaluation de situation; ...

Dilemmes éthiques et conduite autonome

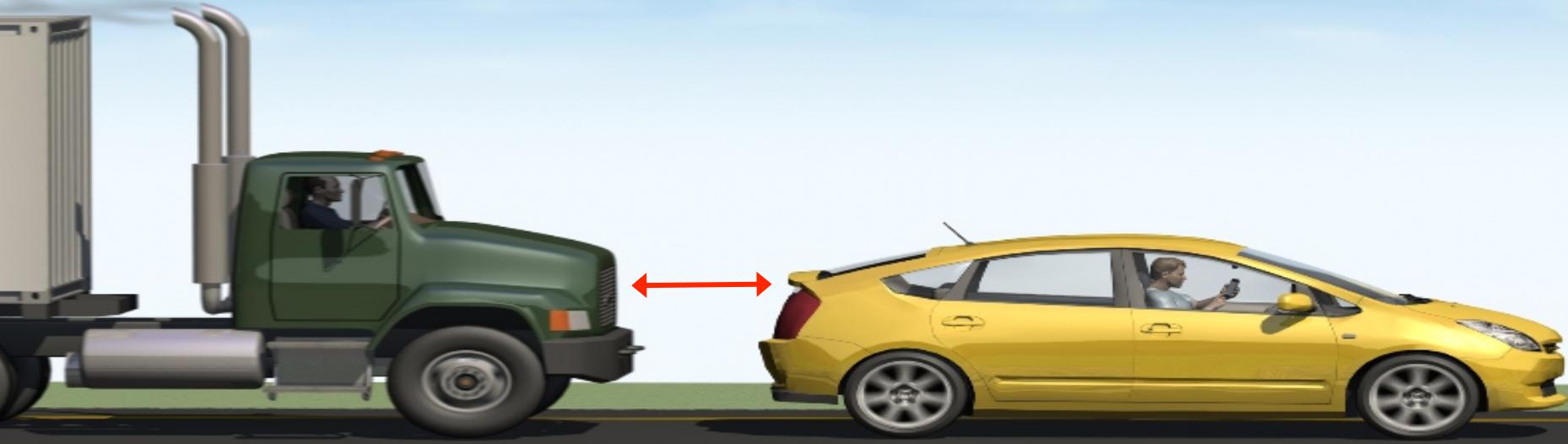


Situations idéalisées (expériences de pensée) permettant de raisonner sur les valeurs et les choix éthiques

Le problème du trolley



Dilemmes éthiques



Situation prévisible



Obstacle imprévu



Situation évolutive; prédition



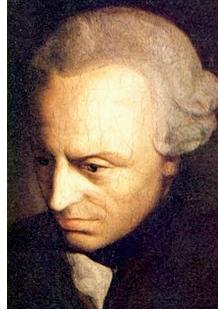
Influence du contexte



Théories éthiques

- Théorie Déontique: respect en toute circonstance d'un impératif moral
- Théorie utilitariste: choix minimisant le conséquences négatives
- Théorie casuistique: détermination au cas par cas
- Théorie Rawlesienne: la justice comme fondement de l'éthique

Ethique Déontique

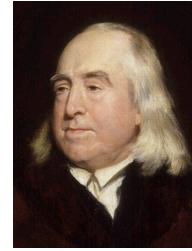


Kant (1797)

- Définir un **impératif moral**, valide en toutes circonstances et situations
- Le non-respect de l'impératif moral est non éthique et rend l'agent de toute conséquence
- Le “devoir” est plus important, plus moral, que le “bien”
- Exemples:
 - “Ne jamais tuer”
 - “Toujours dire la vérité ”
 - “Toujours protéger les humains”?
 - “Toujours protéger le robot”?

Conséquentialisme

Jeremy Bentham (1789), John Stuart Mill (1861)



- Utilitarisme: “Le plus grand bien pour le plus grand nombre”
- Seules les conséquences comptent dans un choix moral
- Les valeurs déterminent le choix de l’action
- Exemples:
 - Tuer une personne plutôt que cinq
 - Tuer un adulte plutôt qu’un enfant

Casuistique et éthique de situation

- Ethique pratique
- Souplesse et pragmatisme
- Casuistique:
 - Règles générales, mais décisions selon le cas.
 - Taxonomies de cas pour élaborer des indications générales
 - Exemple: Bioéthique
- Ethique de situation (Joseph Fletcher)
 - relativisme
 - Les décisions dépendent du contexte et de la situation précise.

Apprendre la morale?



Responsabilité (accountability) des agents autonomes

- Dilution des responsabilités
 - Le concepteur ?
 - Le constructeur ?
 - Le vendeur ?
 - L'opérateur ?
 - L'utilisateur ?
 - L'agent autonome ?
- Création d'une « personnalité morale » pour un agent autonome ?!

Recommandation de la CERNA sur l'Autonomie et capacités décisionnelles

- **Reprises en main**

Le chercheur doit se poser la question des **reprises en main** que l'opérateur ou l'utilisateur peut effectuer (au détriment du robot) et que la machine peut effectuer (au détriment de l'humain), des circonstances qui les permettent ou les rendent obligatoires. Il doit également étudier la possibilité ou non laissée à l'humain de « débrayer » les fonctions autonomes du robot.

- **Limites des programmes**

Le chercheur doit être attentif à **évaluer** les programmes de perception, d'interprétation et de prise de décision et à en **expliciter les limites**. En particulier, les programmes qui visent à conférer une **conduite morale** au robot sont soumis à de telles limites.

Autonomie et capacités décisionnelles

- **Caractérisation de situation**

En ce qui concerne les logiciels d'interprétation du robot, le chercheur doit évaluer jusqu'à quel point ceux-ci peuvent caractériser correctement une **situation** et discriminer entre plusieurs situations qui semblent proches, surtout si la décision d'action prise par l'opérateur ou par le robot lui-même est fondée uniquement sur cette caractérisation. Il faut en particulier évaluer comment les **incertitudes** sont prises en compte.

- **Prévisibilité du système humain-robot**

De manière plus globale, le chercheur doit analyser la **prévisibilité du système humain-robot** considéré dans son ensemble, en prenant en compte les **incertitudes d'interprétation et d'action**, ainsi que les **défaillances** possibles du robot et celles de l'opérateur, et analyser l'ensemble des états atteignables par ce système.

Autonomie et capacités décisionnelles

- **Traçage et explications**

Le chercheur doit intégrer des **outils de traçage** dès la conception du robot. Ces outils doivent permettre d'**élaborer des explications**, même limitées, à plusieurs niveaux selon qu'elles s'adressent à des experts de la robotique, à des opérateurs ou à des utilisateurs.

- **Influences sur le comportement de l'opérateur**

Le chercheur doit être conscient des phénomènes de **biais de confiance**, c'est-à-dire la tendance de l'opérateur à s'en remettre aux décisions du robot, et de **distanciation morale** (« Moral Buffer ») de l'opérateur par rapport aux actions du robot.

- **Décisions à l'insu de l'opérateur**

Le chercheur doit faire en sorte que les **décisions du robot ne soient pas prises à l'insu de l'opérateur** afin de ne pas créer de ruptures dans sa compréhension de la situation (c'est-à-dire afin que l'opérateur ne croie pas que le robot est dans un certain état alors qu'il est dans un autre état).



*“Does your car have any idea why
my car pulled it over?”*

Bio-mimétisme



AIST



Ricky Ma (Hong Kong)



Osaka U.

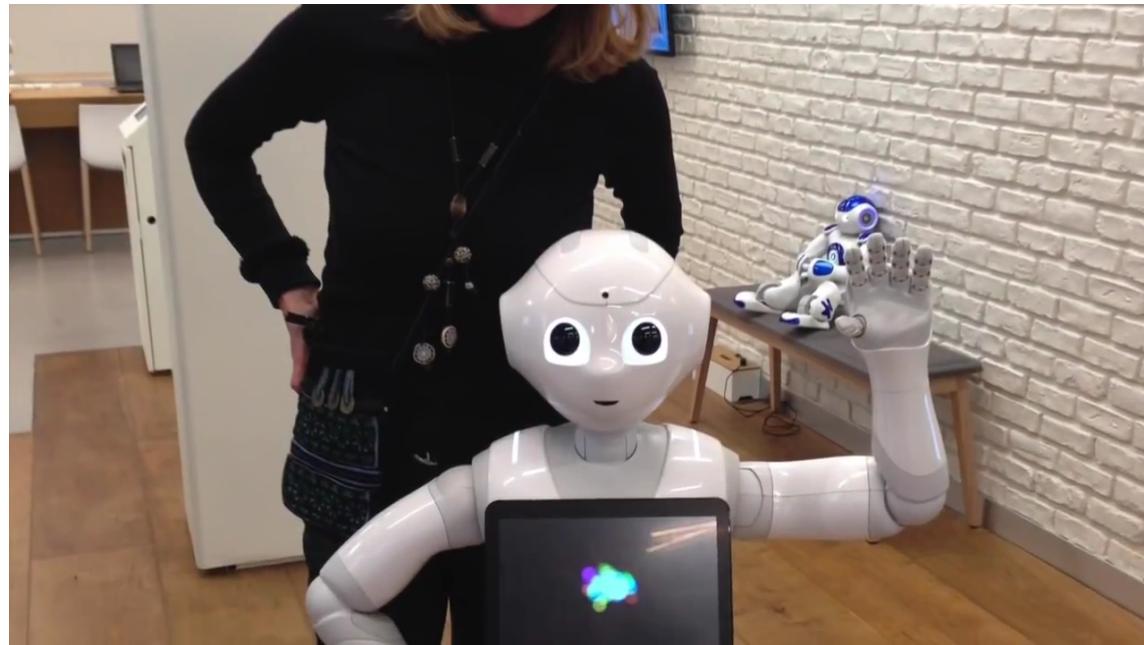




Relations affectives



SOURCE: SoftBank



Recommandations de la CERNA sur l'imitation du vivant et l'interaction affective et sociale avec les humains

- **Frontière vivant - artefact**

Si une **ressemblance quasi parfaite** est visée, le chercheur doit avoir conscience que la démarche biomimétique peut brouiller la frontière entre un être vivant et un artefact. Le chercheur consultera sur ce brouillage le **comité opérationnel d'éthique de son établissement**.

- **Étude des effets**

Pour les projets de recherche qui ont trait au développement de la **robotique affective**, le chercheur s'interrogera sur les répercussions éventuelles de son travail sur les **capacités de socialisation de l'utilisateur**.

- **Interaction enfant-robot**

Pour les projets qui mettent en présence des enfants et des robots, le chercheur doit se poser la question de l'**impact** de l'interaction enfant-robot **sur le développement des capacités émotionnelles de l'enfant**, tout particulièrement dans la petite enfance.

Réhabilitation et augmentation de l'humain

Question sur la nature humaine

EPFL



39

RIC Institute



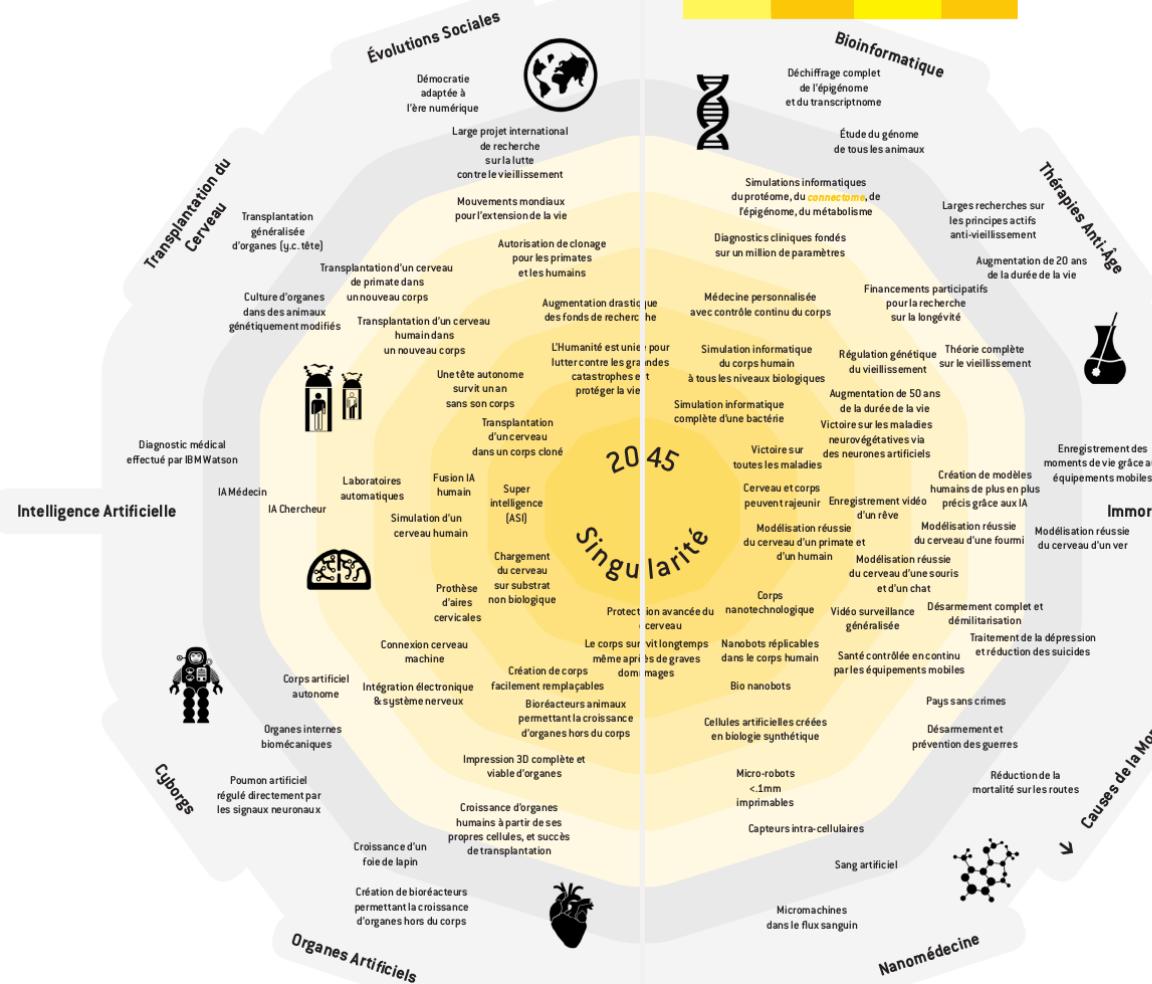
Ekso
Bionics

R. Chatila



RB3D

Singularité et « enhancement »



Recommandations de la CERNA sur la réparation et l'augmentation de l'humain par la machine

- **Autonomie et intégrité de l'individu**

Dans le cas des organes robotisés à vocation **réparatrice**, le chercheur aura le souci de la **préservation de l'autonomie de l'individu** équipé, à savoir de la maîtrise qu'il conservera autant que faire se peut sur ses actions, et de la **conservation de l'intégrité** des fonctions autres que celles concernées par la réparation.

- **Réversibilité de l'augmentation**

Dans le cas des dispositifs robotisés visant l'**augmentation**, le chercheur veillera à la **réversibilité** de celle-ci : les dispositifs doivent être amovibles sans dommage pour la personne, autrement dit, sans que la personne perde l'usage de ses fonctions initiales.

Réparation et augmentation de l'humain par la machine

- **Discrimination induite par l'augmentation**

Le chercheur se posera la question de **l'incidence de l'augmentation** des facultés et des capacités humaines induites par les dispositifs qu'il développe **sur le comportement social** de ceux qui en bénéficient ainsi que, symétriquement, de ceux qui n'en bénéficient pas.

Questions sur l'IA Générale et la superintelligence artificielle

- Pas à l'ordre du jour... mais quand ça arrivera il sera trop tard.
- Mesures préemptives
- Mécanismes d'observation de
- Développement « confiné »

Conclusions

- La science et la technologie impactent la société. Les applications seront difficiles à prédire.
- La robotisation et l'IA peuvent transformer les emplois, en supprimer et en créer.
- Cadrer le développement de la robotique et de l'IA dans une démarche éthique (recherche responsable).
- L'usage des robots et de l'IA amènera à de nouvelles législations (drones, voitures autonomes, armes autonomes, personnes vulnérables,...).



Des initiatives ELS sur le plan international



- <http://www.europarl.europa.eu/committees/en/juri/subject-files.html;jsessionid=BC99F48A4A420741A24E04FD184C34C6.node2?id=20150504CDT00301>
- <http://cerna-ethics-allistene.org/>
- [http://www2.assemblee-nationale.fr/14/les-delegations-comite-et-office-parlementaire/office-parlementaire-d-evaluation-des-choix-scientifiques-et-technologiques/\(block\)/24974](http://www2.assemblee-nationale.fr/14/les-delegations-comite-et-office-parlementaire/office-parlementaire-d-evaluation-des-choix-scientifiques-et-technologiques/(block)/24974)
- <https://www.whitehouse.gov/blog/2016/05/03/preparing-future-artificial-intelligence>
- <http://www.bsigroup.com/en-GB/about-bsi/media-centre/press-releases/2016/april/-Standard--highlighting-the-ethical-hazards-of-robots-is-published/>
- <http://www.japantimes.co.jp/news/2016/04/29/national/japan-pushes-basic-ai-rules-g-7-tech-meeting/>
- http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html