Colloque Humain et Numérique en Interaction
CNRS, 1/02/2019

Intelligence Artificielle et Robotique: Quelle éthique?

Raja Chatila
Institut des Systèmes Intelligents et de Robotique (ISIR)
Sorbonne Université, Paris, France
Raja.Chatila@sorbonne-universite.fr

Membre de la CERNA, Commission de réflexion sur l'éthique de la recherche en sciences et technologies du numérique
Chair, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems

Applications de l'IA et de la robotique: production industrielle, milieux hostiles, santé, services, transport, agriculture, construction, loisirs, défense, …

- Remplacer les humains
(tâches répétitives, dangereuses)



- Assister les humains











- Réhabiliter/augmenter les humains

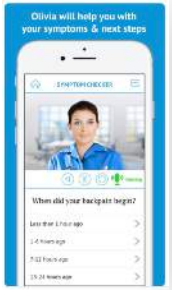


Questions “Ethiques, Légales Sociétales”

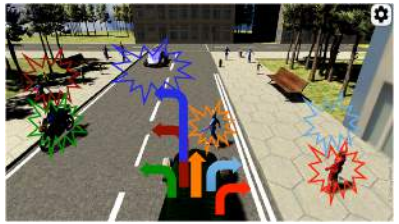
- Transformation du travail
- Vie privée et surveillance
- Données et biais



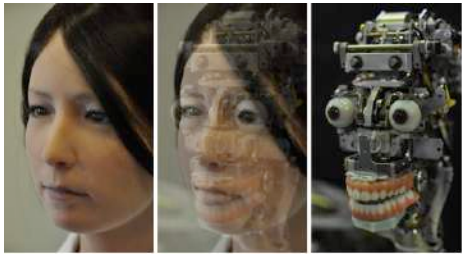

- Transparence, explicabilité des algorithmes
- Décisions autonomes et responsabilité







- Autonomie humaine et manipulation
- Liens affectifs, attachement, isolement
- Dignité humaine
- Transformation de l’être humain



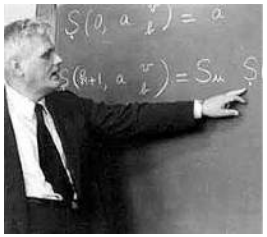
- Statut du robot dans la société

- Usages et applications problématiques (robots sexuels, armes autonomes, …)

HLEG-AI: “Ethical Purpose”

- *“To give an example of the relationship between fundamental rights, principles, and values let us consider the fundamental right conceptualised as ‘respect for **human dignity**’. This right involves recognition of the inherent value of humans (…).*
- *This leads to the ethical principle of autonomy which prescribes that individuals are free to make choices about their own lives, be it about their physical, emotional or mental wellbeing (i.e. since humans are valuable, they should be free to make choices about their own lives).”*

Thèse de Church-Turing, 1936



Alonzo Church

Alan Turing

- Les fonctions **calculables** sont exactement les fonctions récursives (c-à-d représentables par un algorithme qui se termine)
- La Machine Universelle de Turing est un modèle “mécanique” de la calculabilité

Naissance de l'Intelligence Artificielle: Dartmouth College 1956

John McCarthy

Marvin Minsky

The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to **simulate** it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.

Nathaniel Rochester

Claude Shannon

Le robot

- Machine **matérielle** située dans le **monde réel**.
- Un paradigme de l'IA "incarnée" ou "encorporée"
- L'intelligence vue comme l'**interaction rationnelle avec le monde réel**
  - **Perception**; **action** et **mouvement**; **décision.**
  - Communication, **interaction.**
  - **Apprentissage.**
- Capacités développées à divers degrés de complexité, permettant des niveaux d'*autonomie* différents de la machine.

Une définition

- Un système numérique intelligent est nécessairement une machine de Turing Universelle
- Il est basé sur des algorithmes utilisant des données pour résoudre des problèmes [plus ou moins] complexes dans des situations [plus ou moins] complexes.
- Ce système peut inclure la capacité d'améliorer ses performances en classant ou en combinant des données (p.ex., par apprentissage) ou en évaluant des actions antérieures pour en sélectionner de meilleures (p.ex par apprentissage par renforcement).

Les machines peuvent-elles prendre des décisions éthiques?

- L'autonomie humaine est fondée sur l'agentivité - la capacité de délibérer et d'agir intentionnellement et en conscience.
- Les processus computationnels (algorithmes) sont conçus par les humains; les machines opère à un niveau syntactique.
- Les machines n'ont pas une compréhension des fondements de leurs décision et de leurs actions.
- La dignité humaine n'est pas explicitement décrite et ne peut être apprise à une machine syntactique.
- Les machines ne peuvent pas déterminer les valeurs éthiques.

La question de confiance

- Les machines ne peuvent pas prendre des décisions éthiques
- … mais elles peuvent effectuer des choix et des actions qui ont des conséquences éthiques
- Confiance: les machines doivent respecter les valeurs décrites par les humains.

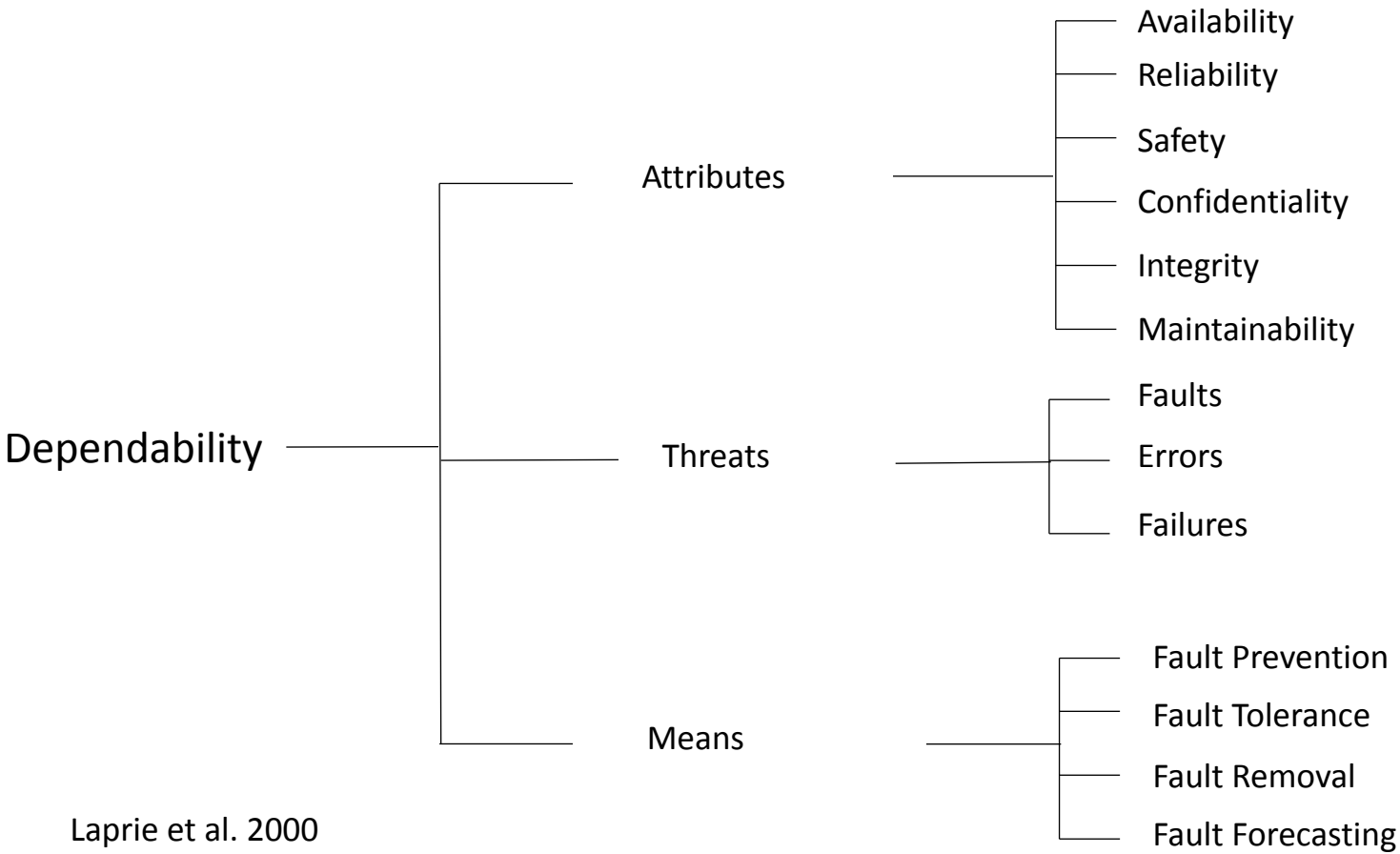
Principes éthiques (AI4People; HLEG-AI)

- Principe de bienfaisance: “faire le bien”
- Principe de non-malfaisance: “ne pas faire le mal”
- Principe d’autonomie: “préserver l’agentivité humaine”
- Principe de justice: “être juste”
- Principe d’explicabilité: “agir de manière transparente”

Principes éthiques pour les systèmes Autonomes et Intelligents
(IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems)

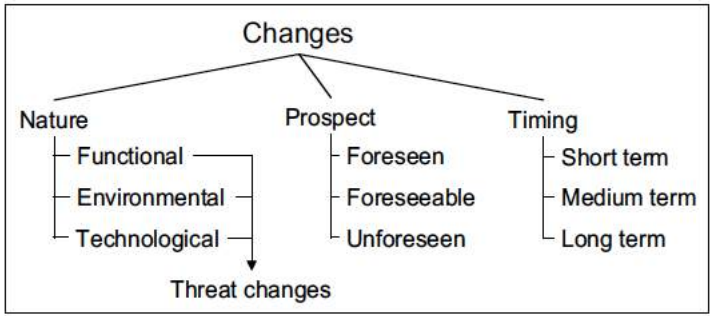
- Respect des droits humains
- Amélioration du bien-être humain
- Autonomie humaine dans la maîtrise des données personnelles, de l'identité numérique
- Aptitude à l'objectif: le système doit être conçu de manière à accomplir la tâche attendue et être prévisible
- Transparence: Il doit toujours être possible de connaître les tenants des décisions
- Responsabilité (Accountability): La responsabilité demeure celle des humains qui sont derrière les systèmes
- Anticiper et prévenir les mes-usages et conséquences inattendues
- Compétence humaine dans la création et l'utilisation des systèmes

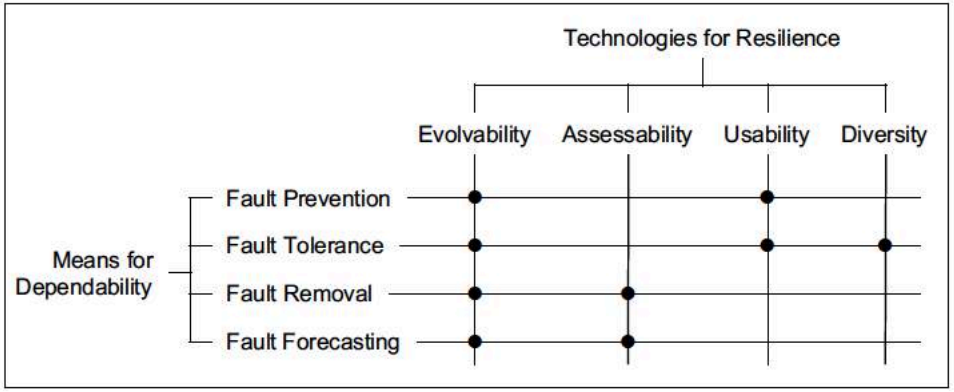
Dependability: Delivery of service that can justifiably be <u>trusted</u>
Resilience: The persistence of service delivery that can justifiably be trusted, when facing changes.



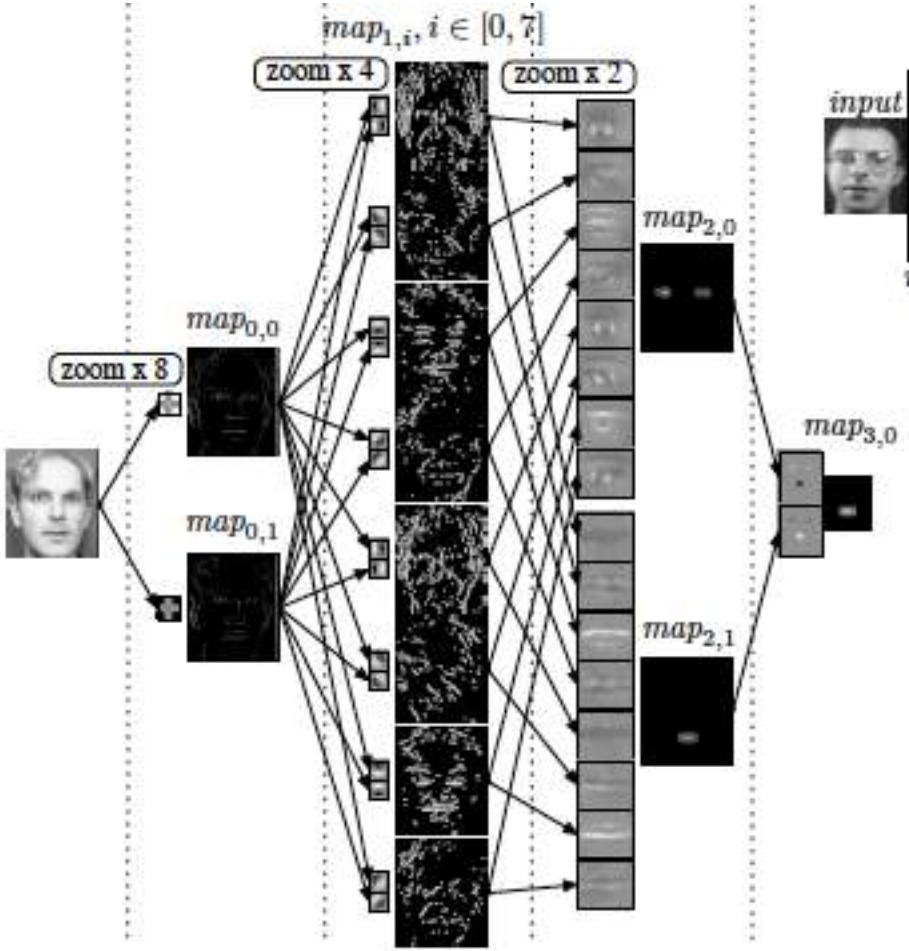
Laprie et al. 2000

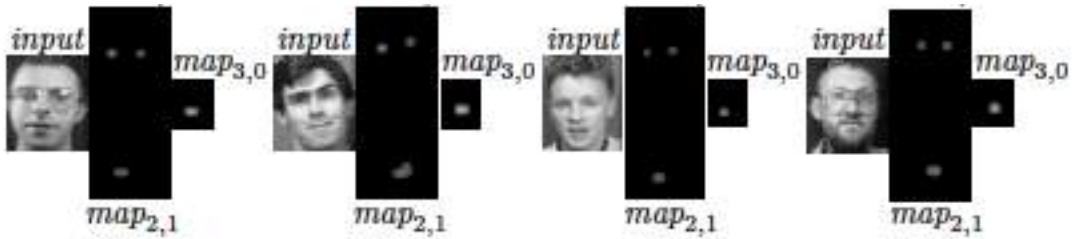
Résilience


Changes
Nature
Functional
Environmental
Technological
Threat changes
Prospect
Foreseen
Foreseeable
Unforeseen
Timing
Short term
Medium term
Long term


Technologies for Resilience
Evolvability
Assessability
Usability
Diversity
Means for Dependability
Fault Prevention
Fault Tolerance
Fault Removal
Fault Forecasting

Apprentissage pour reconnaissance de visages





Biais

- Données représentant la populations?
- Choix des caractéristiques extraites?
- Signification des classes?
- …

Architecture

- Structure et paramètres du réseau?

Confiance de l'utilisateur dans les résultats

- Transparence: qu'est-ce qu'un visage?

Les algorithmes ignorent le contexte

Ningbo, Chine

(image: Weibo)

Implementation

- Ethique par design : Méthodes de développement pour concevoir des systèmes respectant les principes et les valeurs qui en découlent.
- Vérification et validation adaptées au systèmes apprenants
- Définition de standards pour faciliter le développement des systèmes
- Certification: autorités et organismes capables de fournir des moyens et méthodes de test et de validation pour établir la confiance.