# Crowdsourcing for language learning

## Some considerations from deontological or consequentialist ethics

Jean-Yves Antoine

*LI, Université François Rabelais, Tours*

UNIVERSITÉ FRANÇOIS - RABELAIS TOURS

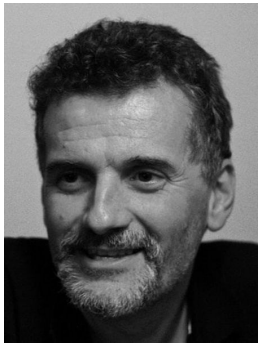
cost

enetCollect, september 7th, 2017

# Some ethical thoughts … from a NLP researcher

**Not an ethicist**

Simply a NLP researcher with ethical concerns



**Natural Language Processing**

- Applications to information retrieval and assistive technologies for disabled people

**Initial questioning : augmentative and alternative communication**


- Word prediction for virtual keyboard
- Severe motion impairments (tetraplegia…)

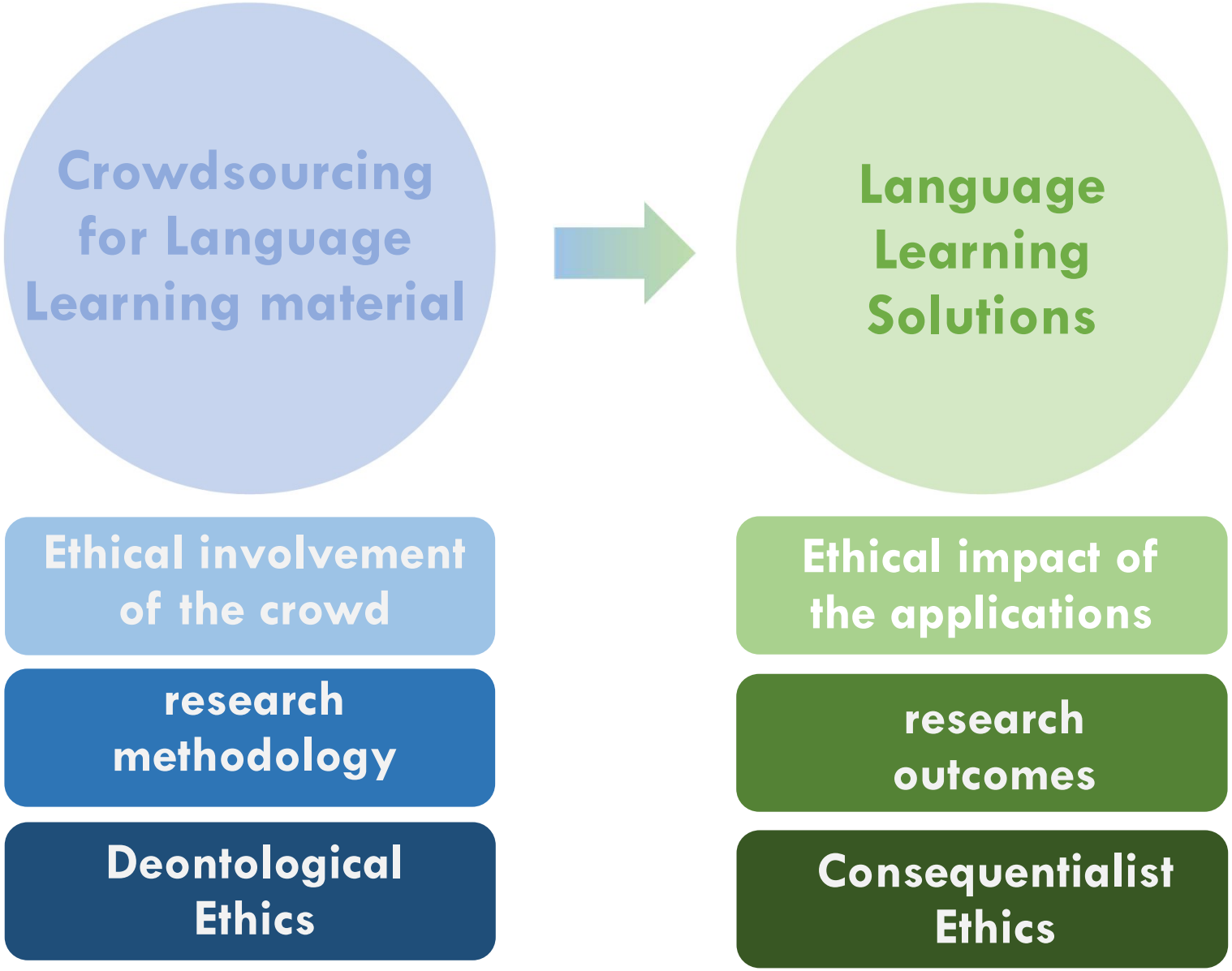
**Increase of input speed**

**Learning of linguistic competencies**

# Ethical researches : aims

- The better intents can have their negative counterparts…
- Systematic assessing the **ethical impact** of new digital technology

# enetCollect



# Ethics and action

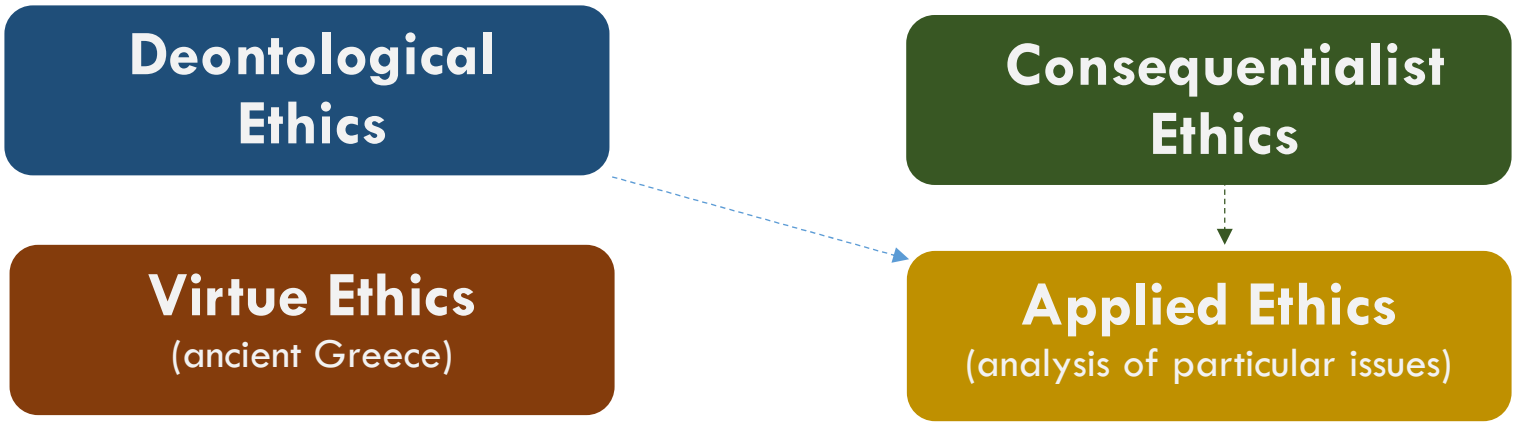
**Law** ***What is allowed and what is forbidden***

- External (state) regulation
- Must be strictly respected

**Ethics** ***What is good/right and what is bad/wrong***

- Duty : self-regulation based on moral judgments
- Elaborated by individuals or some community
- Guide for personal/institutional conduct

**Different approaches of ethics**



# Ethics : approaches

## Deontological Ethics


- **Principle first** – An action must respect some moral principles / rules
- **Ethics** – Debates on the definition of these rules

**Example** John Rawls (1987) A Theory of Justice

→ *Fairness principle*

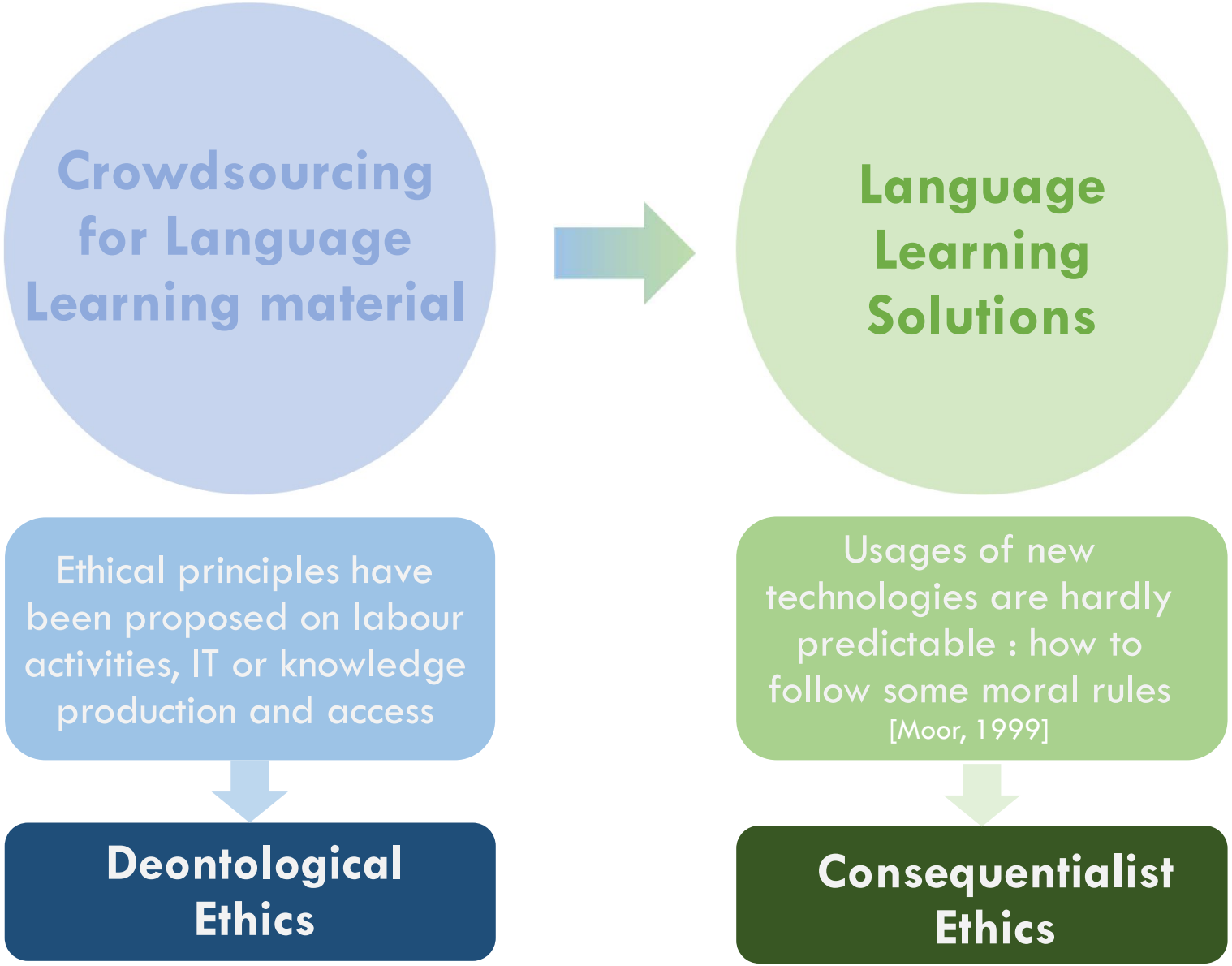
## Consequentialist Ethics

- **Consequence first** – Teleological approach: right actions have good outcomes

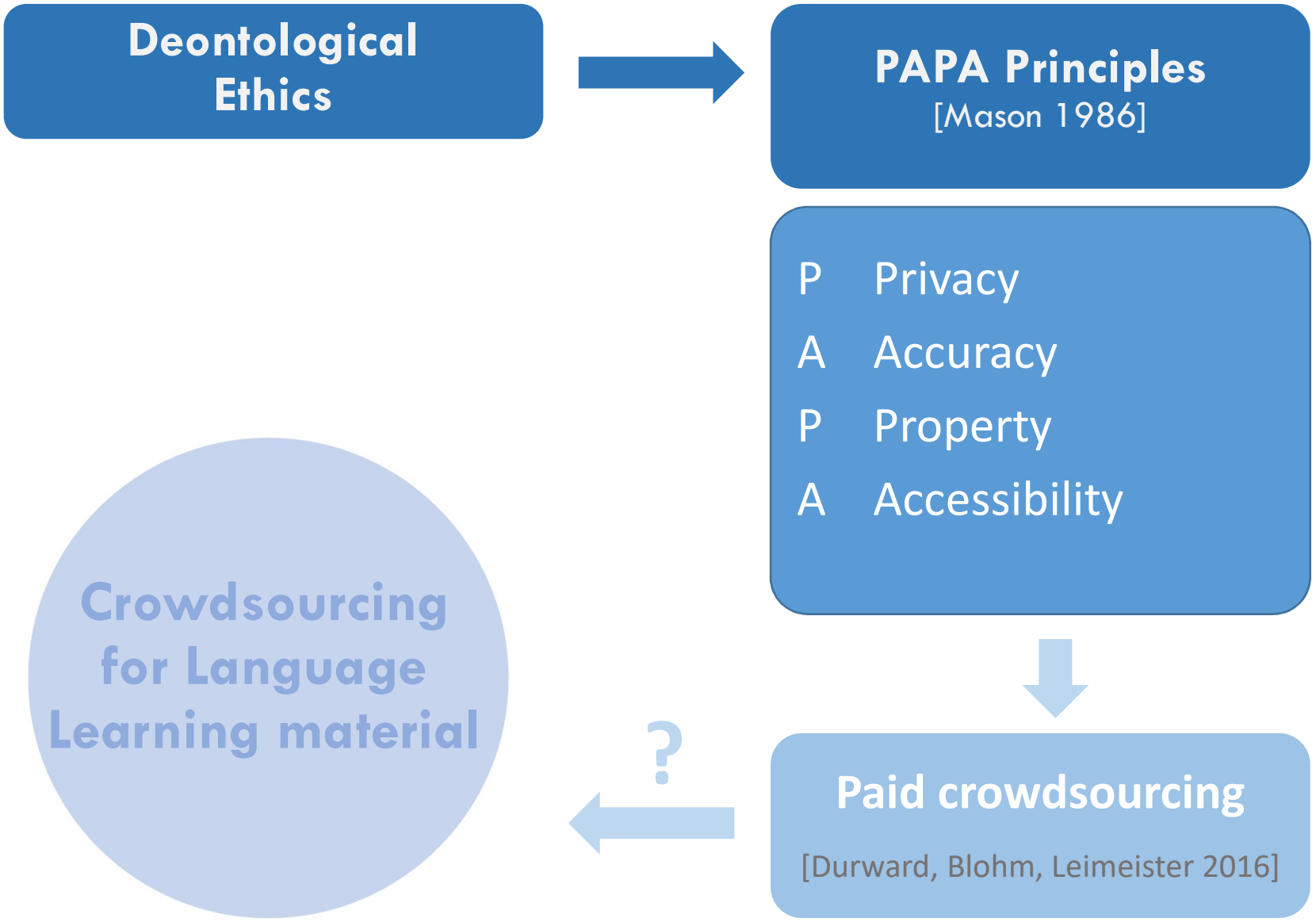
**Examples**

- Utilitarism (Jeremy Bentham, John Stuart Mills)
- Hans Jonas (1990) Imperative of responsibility → *Precautionary principle*

# enetCollect



# Crowdsourcing



# Crowdsourcing : Privacy

**PAPA**
[Mason 1986]

***Ability of the individual to personally control information about oneself***

**Privacy and crowdsourcing –** *Crucial ethical issues on (paid) crowdsourcing platforms, which collect usually a large amount of personal data*

## enetCollect

- **Which personal data ?**
- **Which use and control on these data ?**

- Respect of the regulations on the anonymization / control of data
- European Data Protection Directive 95/46/CE (under revision)

- Privacy go beyond biographical or medical data and personal opinions/preferences
- Individual quality metrics can be seen as an invasion into the privacy of workers [Kajino and al. 2014]

**Question –** *Will performance data be associated to every crowdworker ?*

# Crowdsourcing : Accuracy

**PAPA**
[Mason 1986]

***Extent to which data are correct, reliable and certified*** [Wang & Strong 1996]

**Accuracy and crowdsourcing –** *Central concern of crowdsourcing, but accuracy extends the standard definition of reliability as considered by data science*

enetCollect

**"Pure" Accuracy** (Data Accuracy)

- Training
- Task definition
- User interface
- Reliability metrics

Influence of individual characteristics

- Location (region) ★★★
- Age, book reading frequency ★★
- Openness, conscientiousness ★

[Kazai and al. 2014]

**Privacy**

Must we really look for experts ?

Beyond data accuracy

**Transparency**

# Crowdsourcing : Accuracy

**PAPA**
[Mason 1986]

***Extent to which data are correct, reliable and certified*** [Wang & Strong 1996]

**Accuracy and crowdsourcing –** *Central concern of crowdsourcing, but accuracy extends the standard definition of reliability as considered by data science*

## enetCollect

### Pure accuracy

- Training
- Task definition
- User interface
- Reliability metrics

- Workers profiling

### Transparency [Agerfalk & Fitzgerald 2008]

- **Honesty : explicit vs. implicit (direct vs. indirect)** – Notify workers that a free labour is performed and for which aim

  **Counter example** : reCAPTCHA [Lung 2016]

- **Model transparency** – Workers must be able to understand how their contribution is used in the final data

  **weighted or majority score ? consensus dispute ?** [Doan et al. 2011]

**Question –** *Explicit vs. implicit crowdsourcing ? Which combination model ?*

# Crowdsourcing : Property

**PAPA**
[Mason 1986]

***Intellectual property –*** *Patrimonial and moral property*

**Accuracy and crowdsourcing –** *Collective property, shared between the crowdworkers and the scientists that designed the research and combined the data*

enetCollect

**Patrimonial**

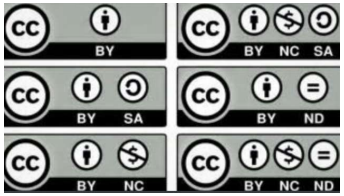
- Right to use ?
- Right to sell ?
- Right to generate incomes ?

**Moral**

- Citation of every crowdworker ?
- Right to modify

- Regulations on intellectual property

**Question –** *Which economical model ?*
*Which copyright / distribution licence ?*



# Crowdsourcing : Accessibility

**PAPA**
[Mason 1986]

***Extent to which data are attainable by the mass of people***
*Technical, economical but also cognitive perspective (data illiteracy, understanding)*

**Accuracy and crowdsourcing –** *Data accessibility but also accessibility to the crowdsourcing procedure : maximising the number and the diversity of participants*

enetCollect

**Data accessibility**

- User-friendly interface, tutorials
- Distribution licence

**Property**

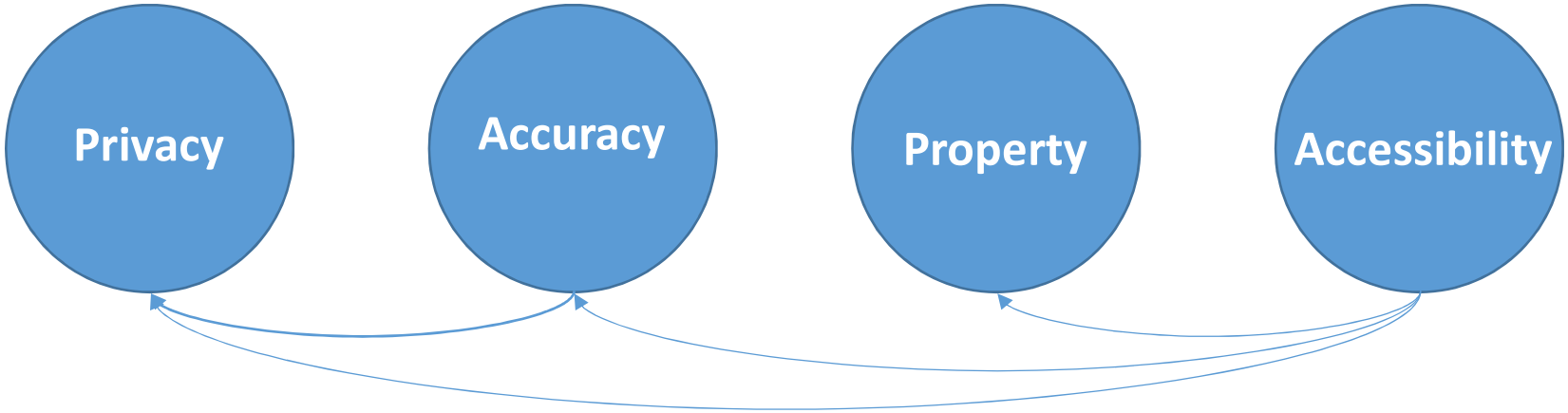
**Privacy + Accuracy**

**Crowdsourcing accessibility**

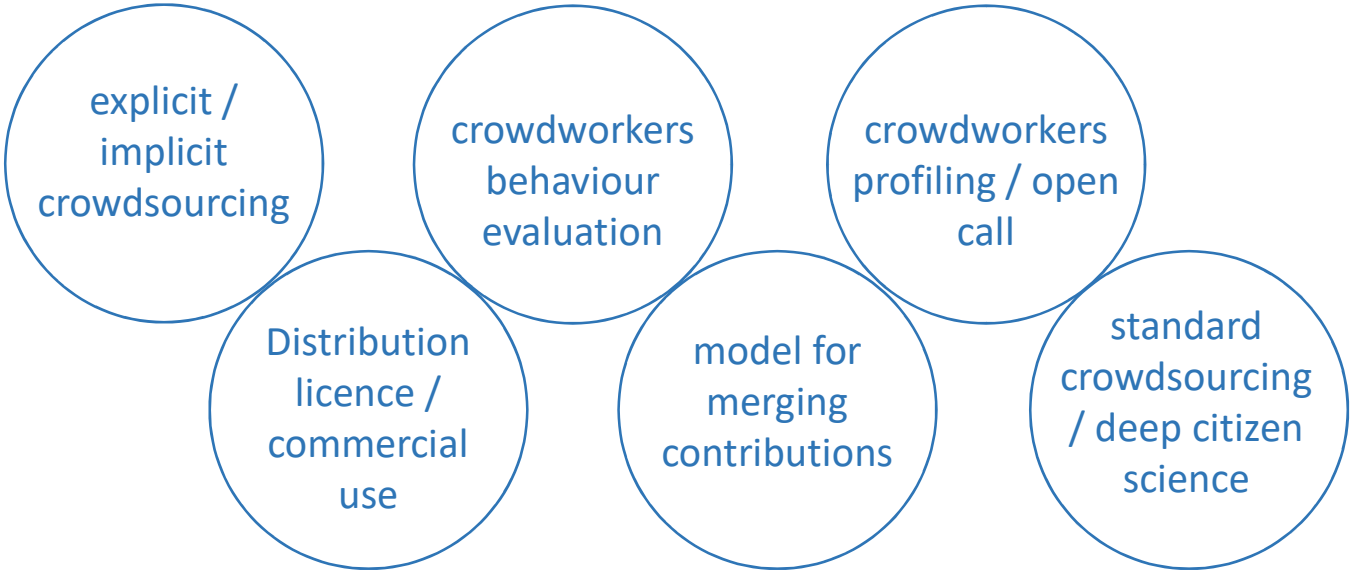
**Equality = maximising the crowd**
- Recruitment strategy
- "no a priori identification" (K. Fort)

**Help the user to contribute**
- Community management, gamification
- User-friendly interface [Doan et al. 2011]
- Recommandation tools [Schnitzer et al. 2015]
- Task adaptation to users' skills

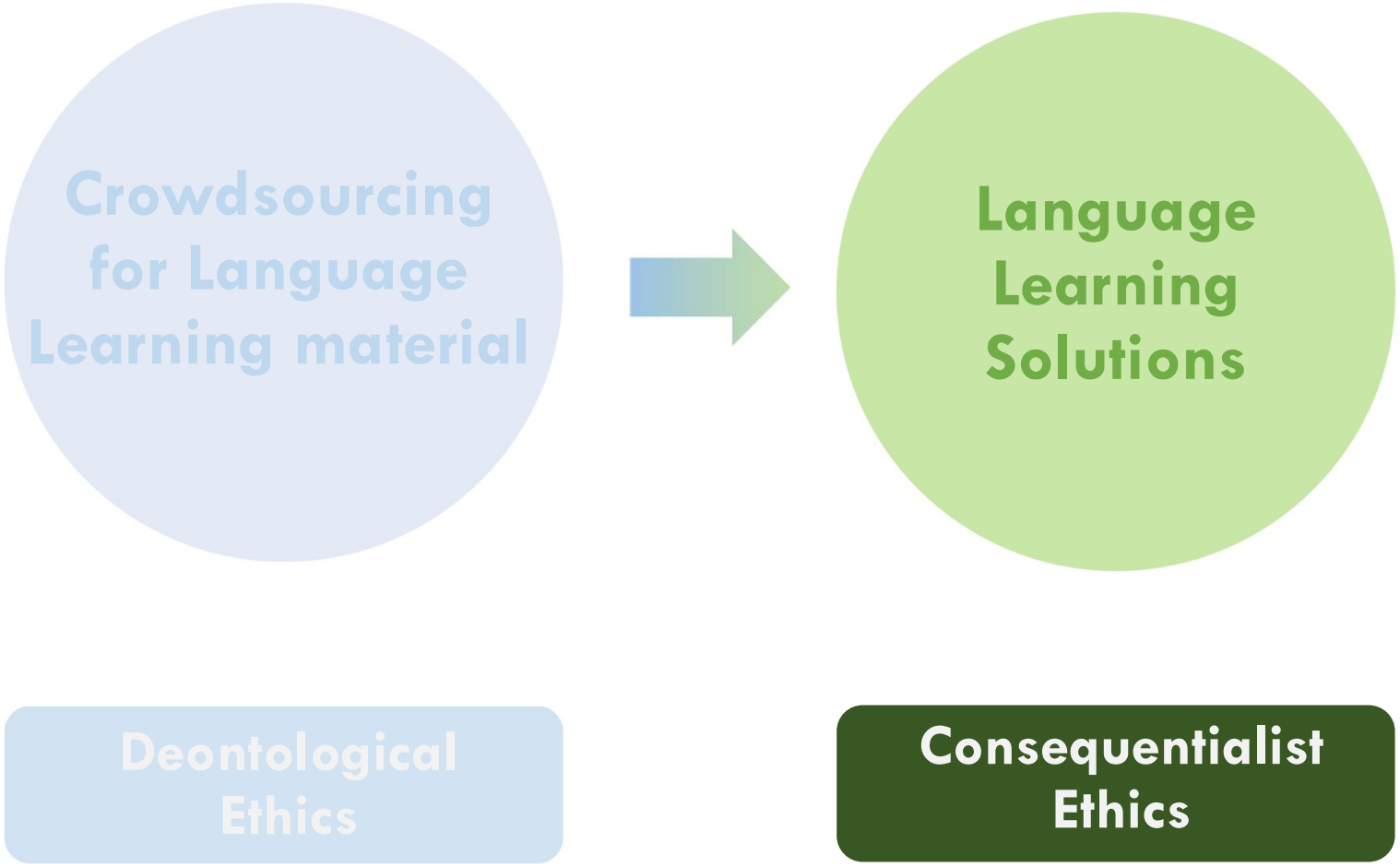
# Crowdsourcing : conclusion




**Some decisions that will impact the ethical value of the project**



# enetCollect



# Language Learning Solutions

**Consequentialist Ethics**

**Consequences of the developed solutions**

- (Industrial) Risk
- Systematic check-up

- Risk ontology
- NLP, HMI & assistive technologies

[Lefeuvre-Halftermeyer et al 2016]

?

**Language Learning Solutions**

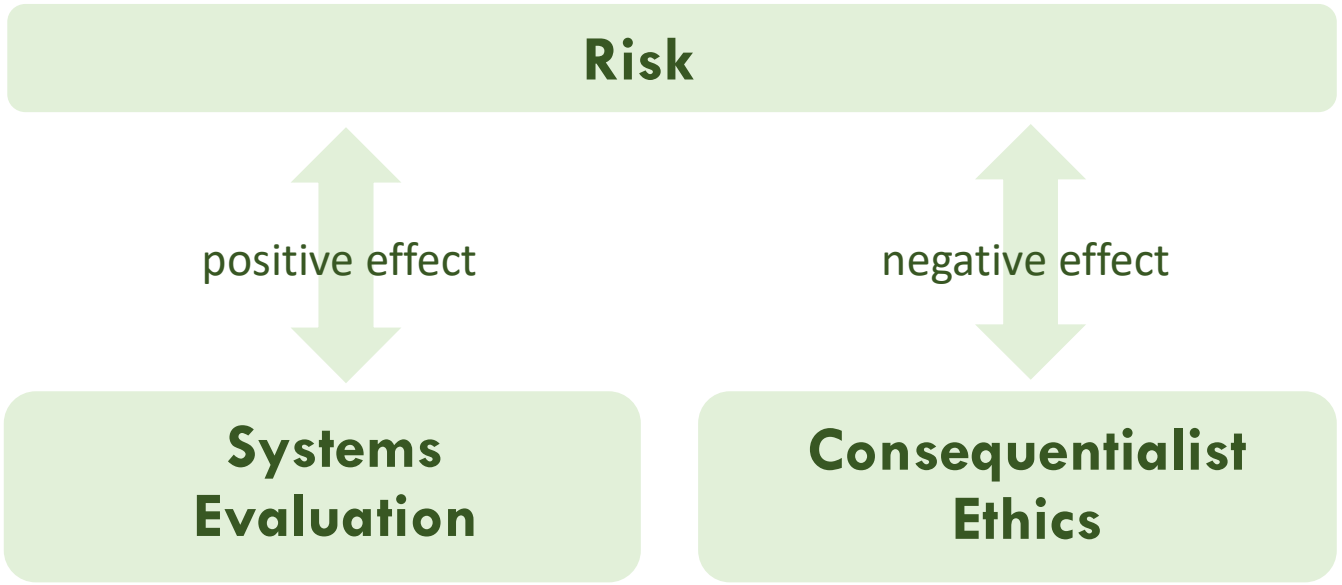
# Risk

**ISO Guide 73** (2009)

**Risk** – Effect of uncertainty on objectives

**Effet** – Deviation from the expected — positive and/or negative

**Objective** – Objectives can have different aspects (such as financial, health and safety, and environmental goals) → *ethical issues*



# Risk analysis

**Risk Source**

Element which alone or in combination has the intrinsic potential to give rise to risk

**Vulnerability**

Something having a susceptibility to a risk source ⇨ Element that should be impacted

**Criticity**

C = LR × L

LR Level of risk (magnitude)

L likehood (probabiliy)





Is the application a risk source ?

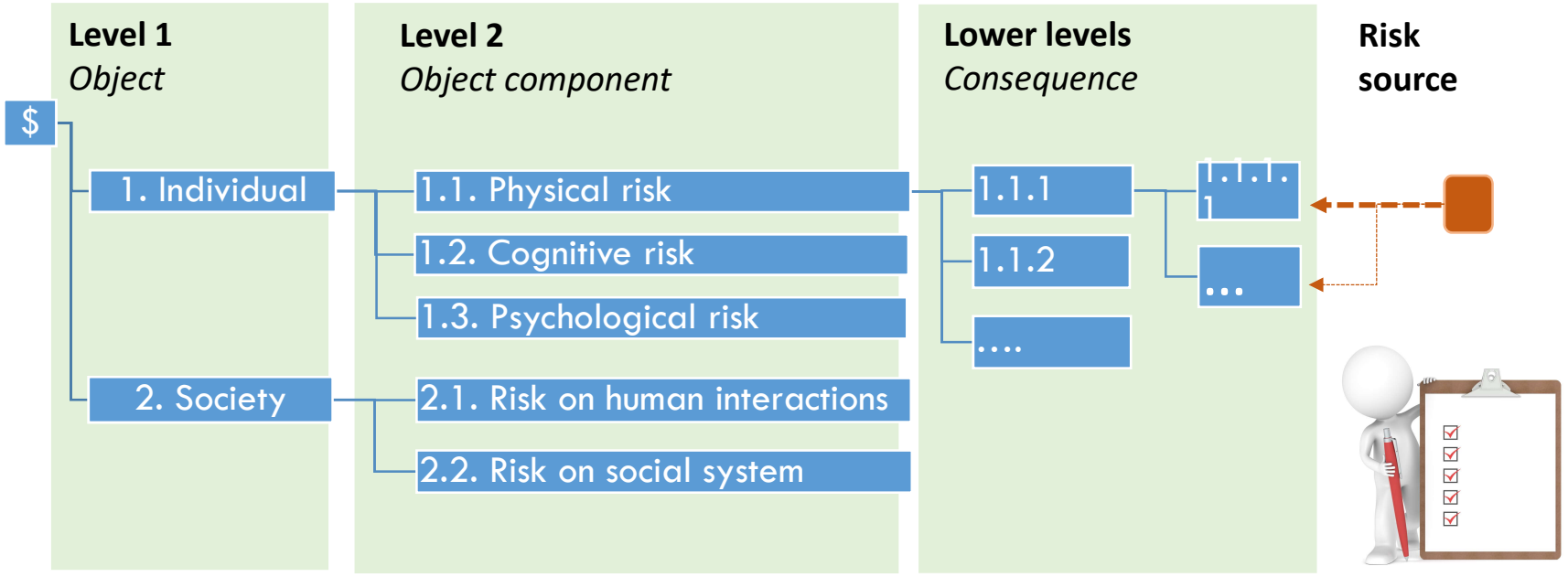
Who / what endures the potential risk ?

enetCollect

**Typology of risk sources**

# Typology of risk sources

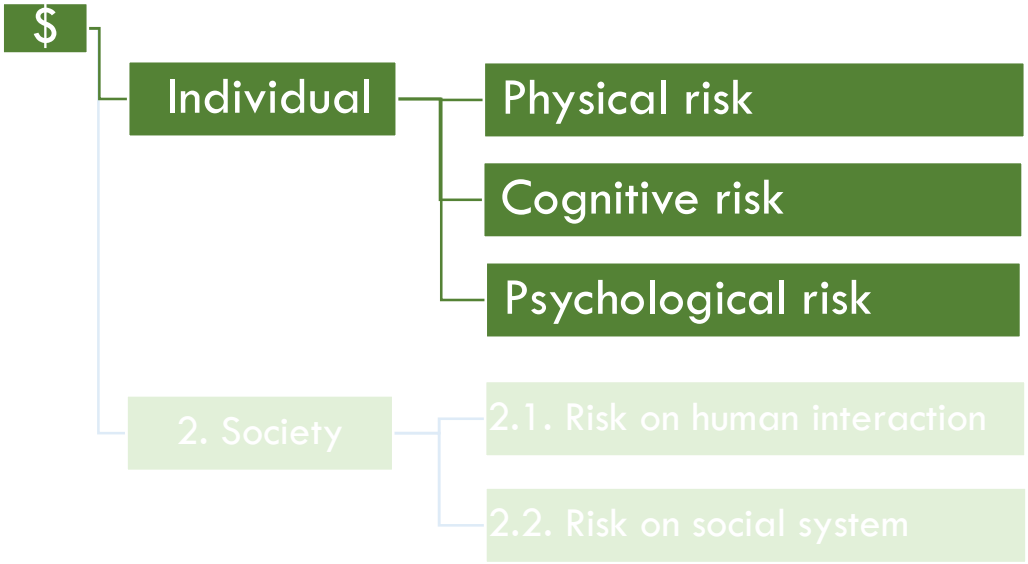
Risk sources organised according to vulnerability




**Risk is multifactorial –** Any risk source should concern several objects of vulnerabilty, while a combination of sources should give rise to a unique risk

⇨ Ontology with graphs rather than a pure hierarchical taxonomy

# Risks on individuals




# Physical risk

**Modification of the physical integrity of an individual or any element**

- Physical risk
  - impact on individual
    - Physical impact
    - Physiological impact
  - impact on physical environnment

**Logicitics : voice control for order pickers**

Productivity increase

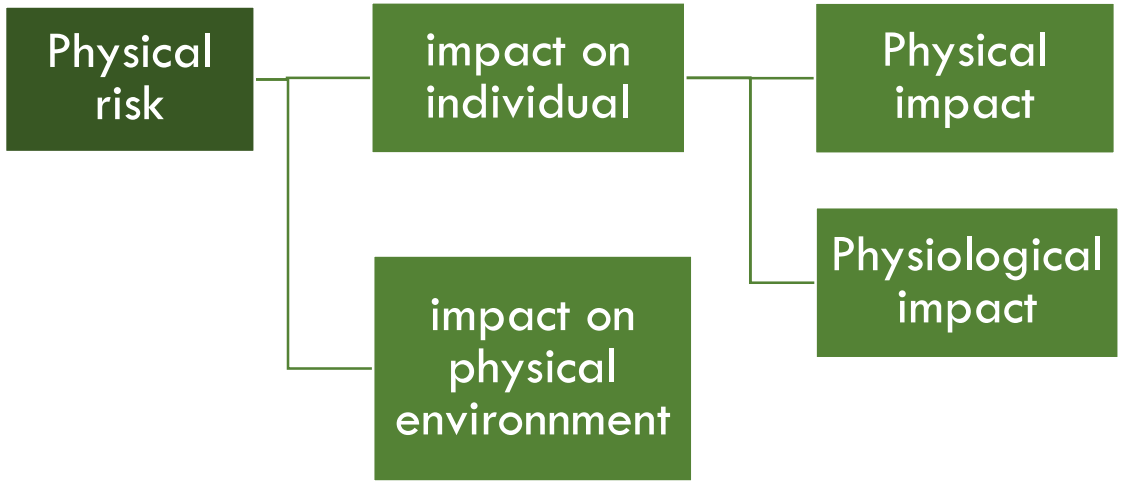
⇨ lombago

⇨ musculo-skeletal disorders

[Govaere and Wioland 2016]

# Physical risk : enetCollect

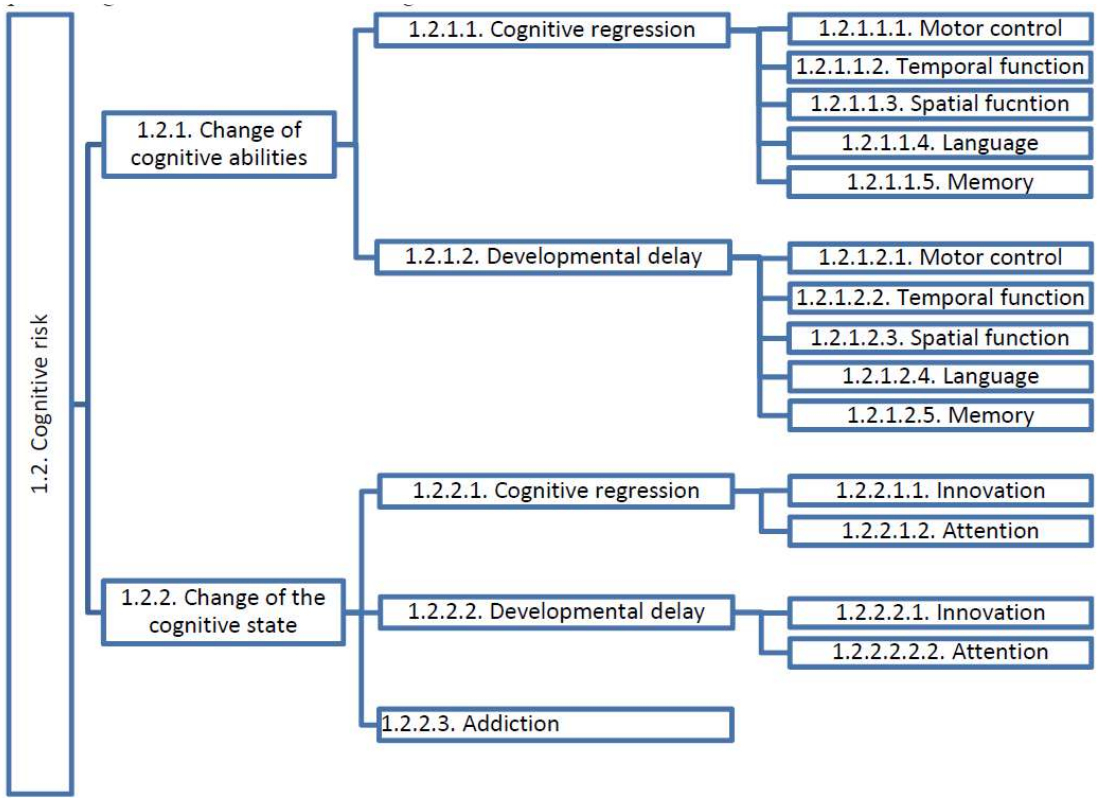
**Modification of the physical integrity of an individual or any element**

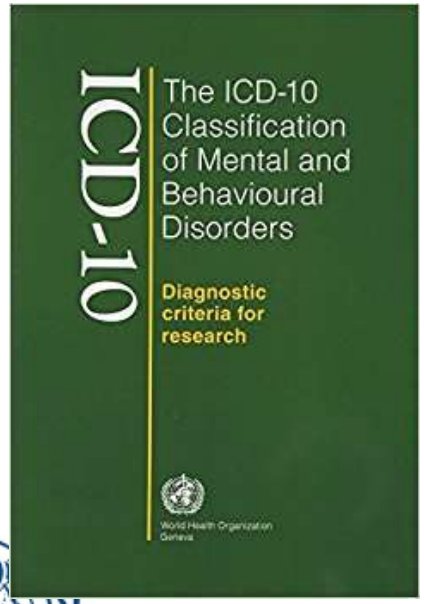



# Cognitive risk

**Modification of a cognitive function or of the general cognitive state**

**Cognitive functions** – mental functions of the ICD-10 (International Classification of Diseases) of the World Health Organisation (WHO/OMS)



# Cognitive risk

**Modification of a cognitive function or of the general cognitive state**

**Information retrieval and cognitive memory : Google *effect*** [Sparrow et al. 2011]

The repetitive use of a search engine leads to sensible modifications of long-term memory

- we are less likely to remember details we believe will be accessible online
- people's ability to learn information offline remains the same

Cognitive risk — Cognitive ability — Cognitive delay — Memory

**Computer-aided writing** – word prediction, automatic translation, orthographic correction…

- Decrease of the cognitive stimulation : impact on language abilities ?
- Long-term tool dependence : lack of autonomy ?

Cognitive risk — Cognitive ability — Cognitive delay — Language

Cognitive regression — Language

# Cognitive risk

**Modification of a cognitive function or of the general cognitive state**

Cognitive risk — Cognitive ability — Cognitive delay — Language

**enetCollect : language abilities**

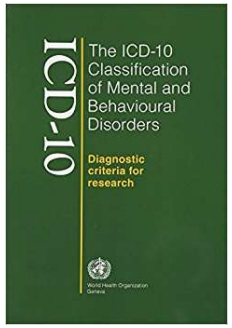
**Language learning applications**

- **Evaluation** – Does the application allow an increase of the linguistic abilities
- **Autonomy** – Do we observe a cognitive tool dependence ?
- **Learning bias** – Does the application favors one ability against others (example : communication skills vs. grammatical mastery)

# Psychological risk

**Temporary or permanent modification of the psychological state**

- Psychological risk
  - Permanent effect
    - Mood disorder
    - Anxiety disorder
    - Personality disorder
  - Temporary effect
    - Psychological fatigue
    - Stress / emotionnal shock

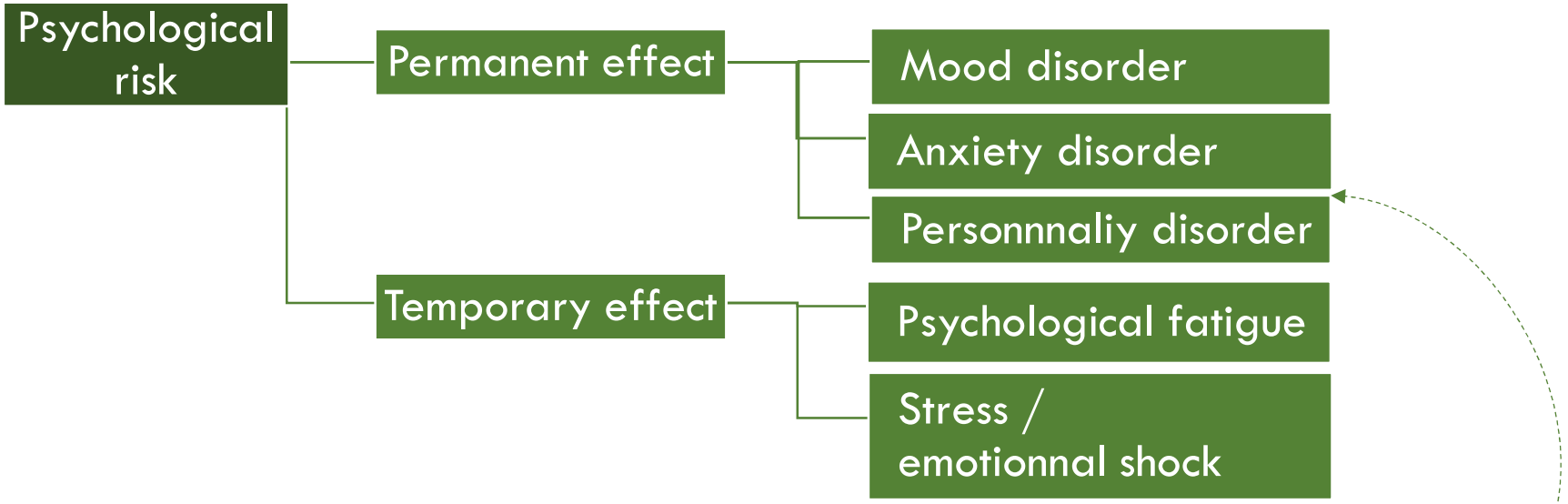



**Speech synthesis for disabled people**

- ⇨ tracheotomy, degenerative disease
- ⇨ New voice, or preventive recording

# Psychological risk

**Temporary or permanent modification of the psychological state**

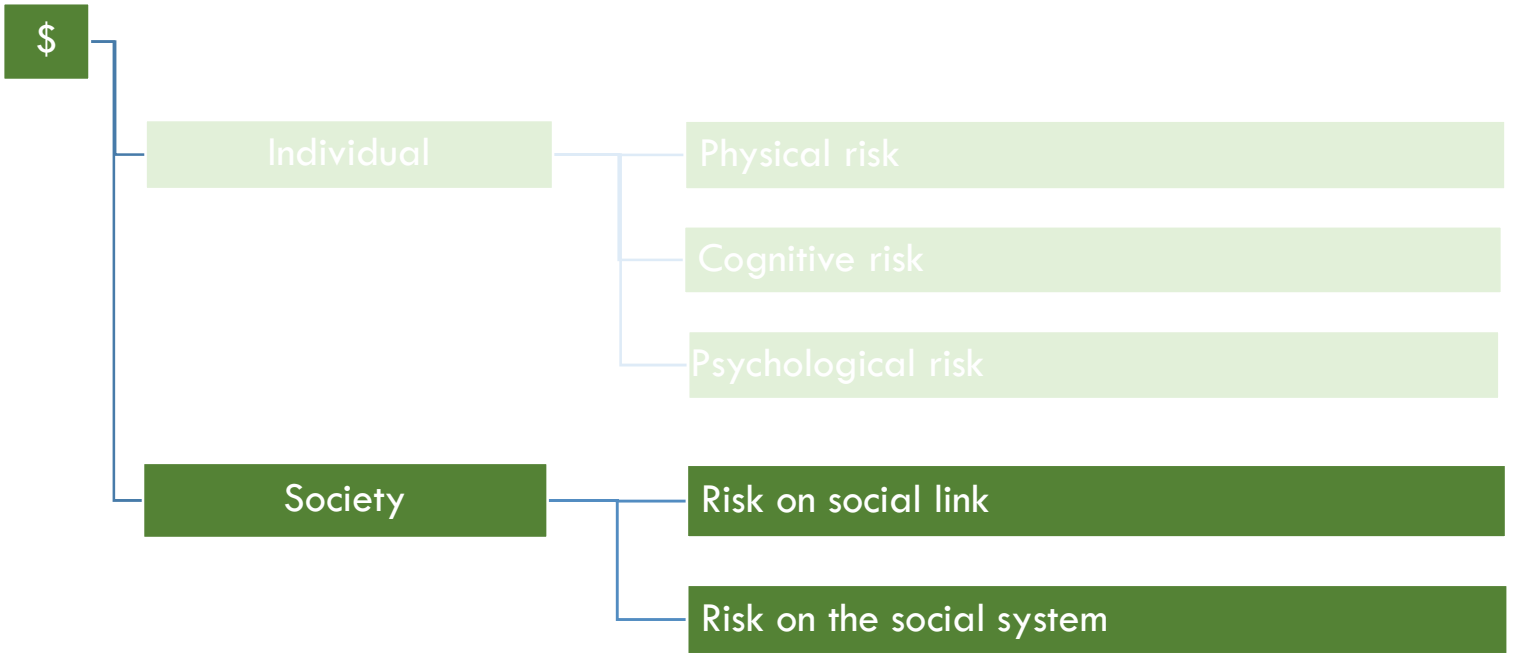



**enetCollect**

**(Self) training**

- ⇨ New technologies can be a source of stress
- ⇨ Training difficulties : feeling of inferiority

# Social risk


$
Individual
Physical risk
Cognitive risk
Psychological risk
Society
Risk on social link
Risk on the social system

# Risk on social link

**Impact on the inter-individuals relations as well as the social insertion**

- Social link
  - Relation with technology
  - Inter-individual relations
  - Individual rights and freedoms
    - Privacy & E-reputation
    - Legal responsibility
    - Labour law
    - Freedom of expression
    - Intellectual property

**Authorship authentication**

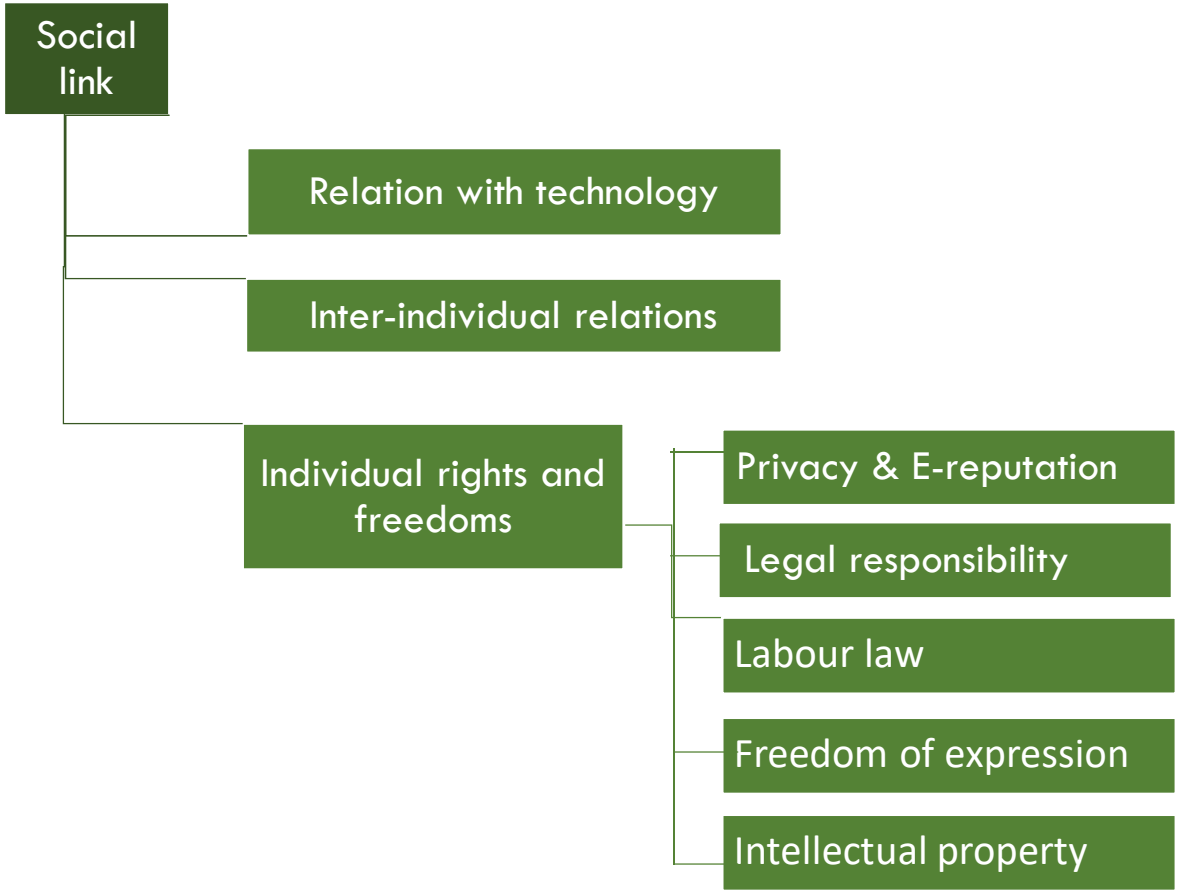
**Crowdsourcing**

Illicit work [Sagot *et al.* 2011]

amazonmechanical turk
Artificial Artificial Intelligence

# Risk on social link

**Impact on the inter-individuals relations as well as the social insertion**



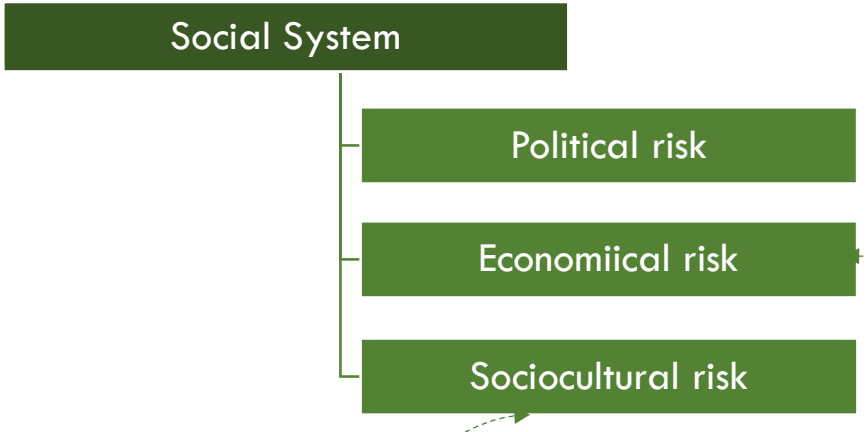




enetCollect

# Risk on the social system

**Modification of the social system :** politics, economy, culture…




**Automatic translation : "*Google*-ese »**

- *English as a pivot language*
- *"natural languages could progressively evolve to seamlessly integrate the linguistic biases of algorithms"* [Kaplan 2014]

**Search engine**

- Paid links, ads: search queries = auction
- (Key)words get an economic value


- Linguistic capitalism: English [Kaplan 2014]

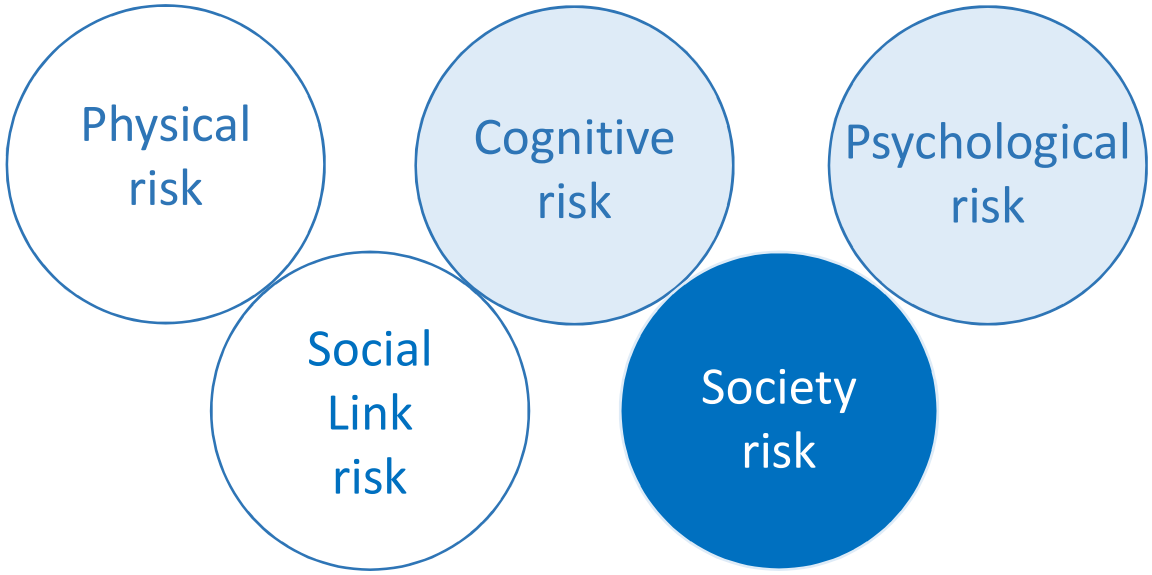
# Risk on the social system

**Modification of the social system** : politics, economy, culture…


enetCollect

# eLearning solutions : conclusion

**Some concerns on individuals as well as society**



Jean-Yves Antoine
*LI, Université François Rabelais de Tours, Blois*

*Jean-Yves.Antoine@univ-tours.fr*

## Multidisciplinary work

Anaïs Lefeuvre-Halftermeyer (NLP)
*LIFO, U. Orléans*

Willy Allègre (Assistive Technologies & occuptational therapy)
*CMRFF de Kerpape, Ploemeur*

Virginie Govaere (ergonomics, occuptational risk)
*LIFO, INSA Bourges*

# References : deontological approach


P.J. Ågerfalk, B. Fitzgerald (2008) Outsourcing to an unknown workforce : exploring opensourcing as a global sourcing strategy. Mis Quaterly. pp. 385-409.

A. Doan, et al. (2011) "Crowdsourcing systems on the world-wide web,"

Communications of the ACM, vol. 54, pp. 86-96, 2011.

D. Durward, I. Blohm, J.M. Leimeister (1986) Is There PAPA in Crowd Work? A Literature Review on Ethical Dimensions in Crowdsourcing. Proc. 2016 *Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress*

H. Kajino, et al. (2014) Preserving worker privacy in crowdsourcing, *Data Mining and Knowledge* Discovery, pp. 1-22.

J. Lung (2016) Ethical and legal considerations of Recaptcha.

R.O. Mason (1986) Four ethical issues of the information age, *Mis Quarterly*, pp. 5-12.

J. Rawls (1971) A Theory of Justice, *Belknap Press of Harvard University*, Cambridge, Mass

S. Schnitzer, C. Rensing, S. Schmidt, K. Borchert, M. Hirt, P. Tran-Gia (2015) Demands on task recommandation in crowdsourcing platforms – the worker's perspective

Y. Wang, Strong (2013) Respecting User Privacy in Mobile Crowdsourcing, *Science*, vol. 2, pp. 50-64.

# References : consequentialist approach


G. Adda, B. Sagot, K. Fort, J. Mariani (2011) Crowdsourcing for Language Resource Development: Critical Analysis of Amazon Mechanical Turk Overpowering Use. Proc. *5th Language and Technology Conference*. Poznan, Poland. 2011

V. Govaere, L. Wioland (2016) Degradations and Consequences of ICT in Occupational Prevention Terms as Illustrated by the Transport and Logistics Sector. Proc. *HCI International'2016*, Toronto, Canada. *Lecture Notes in Computer Science*, *LNCS*, Volume 9752

H. Jonas (1984) The Imperative of Responsibility: In *Search of Ethics for the Technological Age*. University of Chicago Press.

F. Kaplan (2014) Linguistic Capitalism and Algorithmic Mediation, *Representations*, 127 (1), p. 57–63.

A. Lefeuvre-Halftermeyer , V. Govaere , J.Y. Antoine , W. Allegre , S. Pouplin , J-P. Departe , S. Slimani , A. Spagnulo (2016) Typologie des risques pour une analyse éthique de l'impact des technologies du TAL. *Traitement Automatique des Langues*, *TAL*, *57*(2); pp. *47-71*.

B. Sparrow, J. Liu, D.M. Wegner (2011) Google effects on memory: cognitive consequences of having information at our fingertips, *Science*, 333, p. *776-778*.