Le contrôle humain des systèmes algorithmiques

Les Lundis de l'IA et de la Finance
Winston Maxwell
6 mars 2023

TELECOM Paris

INSTITUT POLYTECHNIQUE DE PARIS

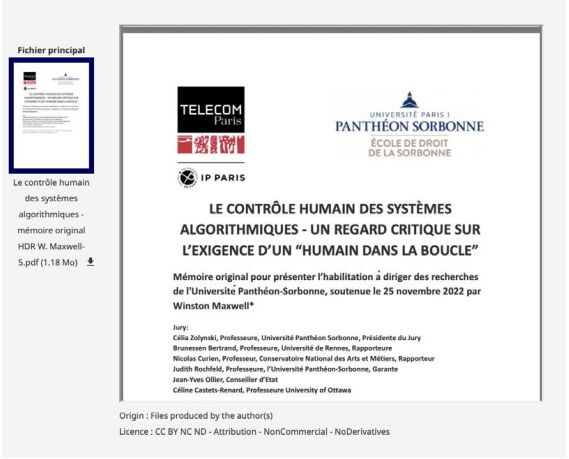
anr agence nationale de la recherche

Projet ANR-20-CHIA-0023
Explainable AI for Anti-Money Laundering

Comment s’assurer que l’humain ajoute de la valeur ?

(et non le contraire…)

HAL : **tel-04010389**



Fichier principal

Le contrôle humain des systèmes algorithmiques - mémoire original HDR W. Maxwell-5.pdf (1.18 Mo)

TELECOM Paris

IP PARIS

UNIVERSITÉ PARIS 1 PANTHÉON SORBONNE
ÉCOLE DE DROIT DE LA SORBONNE

LE CONTRÔLE HUMAIN DES SYSTÈMES ALGORITHMIQUES - UN REGARD CRITIQUE SUR L’EXIGENCE D’UN “HUMAIN DANS LA BOUCLE”

Mémoire original pour présenter l’habilitation à diriger des recherches de l’Université Panthéon-Sorbonne, soutenue le 25 novembre 2022 par Winston Maxwell*

Jury:
Célia Zolynski, Professeure, Université Panthéon Sorbonne, Présidente du Jury
Brunessen Bertrand, Professeure, Université de Rennes, Rapporteure
Nicolas Curien, Professeur, Conservatoire National des Arts et Métiers, Rapporteur
Judith Rochfeld, Professeure, l’Université Panthéon-Sorbonne, Garante
Jean-Yves Ollier, Conseiller d’Etat
Céline Castets-Renard, Professeure University of Ottawa

Origin : Files produced by the author(s)
Licence : CC BY NC ND - Attribution - NonCommercial - NoDerivatives

Plan

1. Le contrôle humain dans les textes
2. Une classification des contrôles humains
3. Les finalités du contrôle humain
4. Les obstacles à un contrôle humain efficace
5. Améliorer le contrôle humain

1. Le contrôle humain dans les textes
2. Une classification des contrôles humains
3. Les finalités du contrôle humain
4. Les obstacles à un contrôle humain efficace
5. Améliorer le contrôle humain

Textes européens: un vocabulaire chaotique

contrôle humain (human oversight)

réexamen individuel par des moyens non-automatisés

human-on-the-loop (HOTL)

“véritable” contrôle humain

contrôle humain effectif

examen humain

intervention humaine

contrôle de décision significatif

garantie humaine

surveillance et vérification humaines

supervision humaine complète à tout moment

maîtrise par l’utilisateur

human-in-the-loop (HITL)

Trois conditions pour un contrôle humain efficace

1. avoir une connaissance du fonctionnement de l'algorithme et de ses limitations

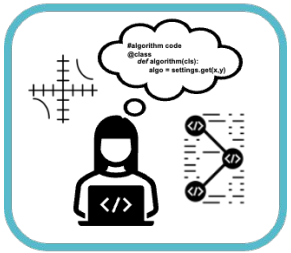
2. s'engager dans un processus cognitif de réflexion qui tient compte du contexte de la décision

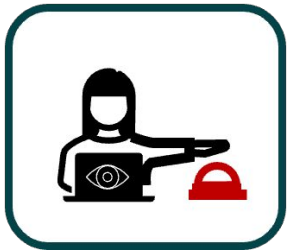
3. avoir l'autorité et la capacité matérielle d'intervenir dans le système et changer la décision

1. Le contrôle humain dans les textes
2. Une classification des contrôles humains
3. Les finalités du contrôle humain
4. Les obstacles à un contrôle humain efficace
5. Améliorer le contrôle humain

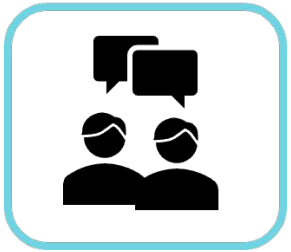
Les contrôles “système” ou “individuels”

Les contrôles “système”

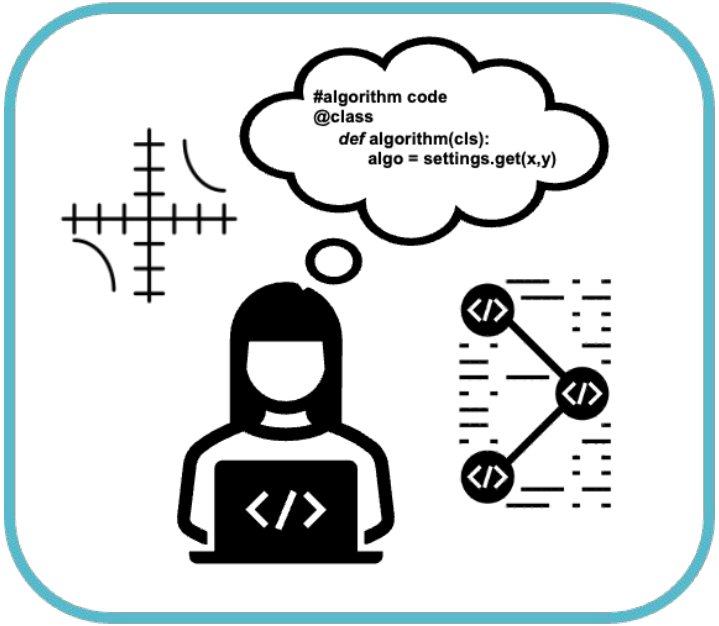




Les vérifications “individuelles”



Contrôle “système” 1: au stade de la conception



CJUE: “modèles et critères préétablis spécifiques”

Loi du 6 janvier 1978: “assurer la maîtrise du traitement algorithmique et de ses évolutions”

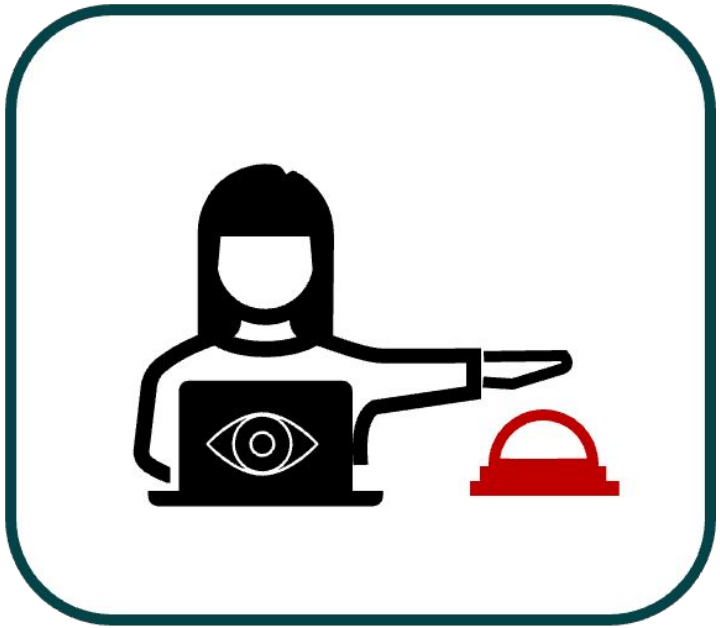
L’humain est responsable du choix des paramètres et de leur pondération

Contrôle “système” 2: au stade des tests et audits

CJUE: “vérifications à intervalles réguliers”

L’humain est responsable de vérifier régulièrement la pertinence du modèle et l’absence de biais

Contrôle “système” 3: Human-on-the-loop (HOTL)



AI Act: “être en mesure de surveiller correctement son fonctionnement, afin de pouvoir détecter et traiter dès que possible les signes d’anomalies, de dysfonctionnements et de performances inattendues”

L’humain doit pouvoir interrompre le fonctionnement en cas d’anomalie

Contrôle “individuel” 1: Human-in-the-loop (HITL)

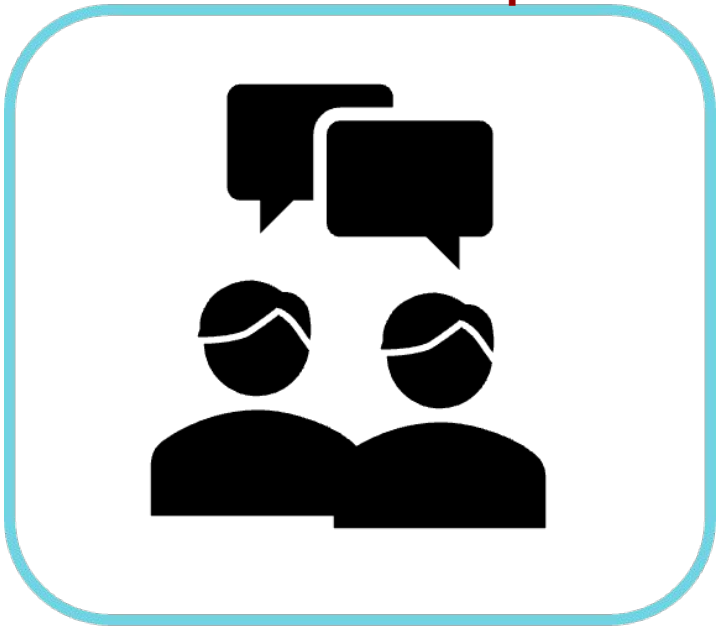
La vérification humaine s’effectue avant la prise de décision.

Trois approches :

1. approche “tribunal”
2. vérification humaine à la lumière d’autres informations
3. vérification humaine SANS autres informations

Contrôle “individuel” 2: au stade des contestations

L’individu conteste une décision algorithmique a posteriori :



RGPD: “exprimer son point de vue”

RGPD, DSA, etc.: “contester”

Projet de directive travailleurs de plateforme: “clarifier les faits, les circonstances, et les raisons qui ont conduit à la décision”

1. Le contrôle humain dans les textes
2. Une classification des contrôles humains
3. Les finalités du contrôle humain
4. Les obstacles à un contrôle humain efficace
5. Améliorer le contrôle humain

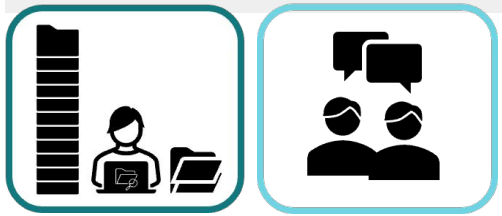
Les trois objectifs du contrôle humain

Détecter les erreurs

- biais et erreurs aléatoires
- détecter les discriminations
- vérifier la performance prédictive

Procédure équitable

- Dignité humaine
- Etre jugé par un autre humain
- Participer à la décision
- “Réversibilité des rôles”
- Due process



Responsabilité

- Fixer la responsabilité
- Justifier une décision
- Responsibility gap
- Démontrer sa conformité

1. Le contrôle humain dans les textes
2. Une classification des contrôles humains
3. Les finalités du contrôle humain
4. Les obstacles à un contrôle humain efficace
5. Améliorer le contrôle humain

Les obstacles à un contrôle humain efficace

Le paradoxe de la performance

Biais de l'automatisation

- complaisance, ennui
- limites cognitives
- pression du temps

Biais de responsabilité

Confusion des tâches

L'explicabilité à double tranchant

L'accès à des informations supplémentaires

1. Le contrôle humain dans les textes
2. Une classification des contrôles humains
3. Les finalités du contrôle humain
4. Les obstacles à un contrôle humain efficace
5. Améliorer le contrôle humain

Trois recommandations

1.Distinguer les finalités du contrôle humain, et adapter les modalités (et KPIs) du contrôle humain à chaque finalité

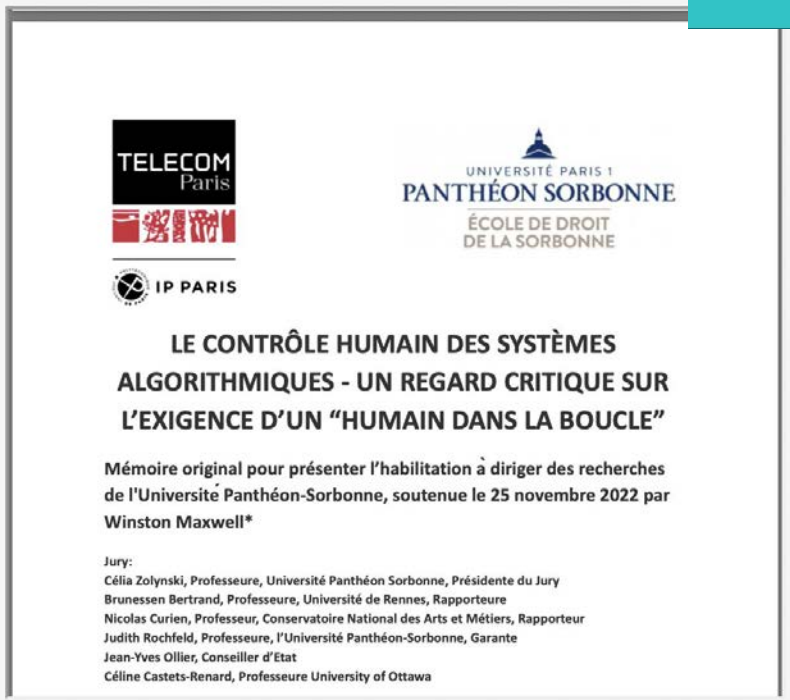
2. Séparer les tâches et les responsabilités
   - entre l'humain et la machine
   - entre les différentes équipes et modalités de contrôle humain

3.Tester l'équipe humain/machine par rapport aux KPIs liés à chaque finalité (erreurs, procédure équitable, responsabilité)

Pour plus d'informations…

Fichier principal

Le contrôle humain des systèmes algorithmiques - mémoire original HDR W. Maxwell-5.pdf (1.18 Mo)



TELECOM Paris
IP PARIS

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE
ÉCOLE DE DROIT DE LA SORBONNE

LE CONTRÔLE HUMAIN DES SYSTÈMES ALGORITHMIQUES - UN REGARD CRITIQUE SUR L'EXIGENCE D'UN "HUMAIN DANS LA BOUCLE"

Mémoire original pour présenter l'habilitation à diriger des recherches de l'Université Panthéon-Sorbonne, soutenue le 25 novembre 2022 par Winston Maxwell*

Jury:
Célia Zolynski, Professeure, Université Panthéon Sorbonne, Présidente du Jury
Brunessen Bertrand, Professeure, Université de Rennes, Rapporteure
Nicolas Curien, Professeur, Conservatoire National des Arts et Métiers, Rapporteur
Judith Rochfeld, Professeure, l'Université Panthéon-Sorbonne, Garante
Jean-Yves Ollier, Conseiller d'Etat
Céline Castets-Renard, Professeure University of Ottawa

Origin : Files produced by the author(s)

HAL : **tel-04010389**

MERCI !