Audition devant la Commission nationale consultative des droits de l'homme (CNCDH)

Comment assurer l'efficacité du contrôle humain dans les systèmes de décision algorithmiques?

12 janvier 2022

Winston Maxwell

Telecom Paris, Institut Polytechnique de Paris
winston.maxwell@telecom-paris.fr

telecom-paris.fr/ai-ethics

Le contrôle humain dans les systèmes de décision algorithmiques

Que signifie “contrôle humain”?

A quoi sert-il?

Comment maximiser son efficacité?

Contrôle humain, une exigence légale

Europe	Etats-Unis	Autres traités internationaux
La Charte des droits fondamentaux (CJUE Avis 1/15 PNR, Affaire C-511/18 La Quadrature du Net)	Due process	Droit international humanitaire (armes autonomes)
La Constitution française (Conseil const. n° 2003-467, 2018-765, 2020-834)	Loi de l'Etat de Washington sur la reconnaissance faciale	Convention de Vienne sur le transport routier (voitures autonomes)
RGPD art. 22 Convention 108+ art. 9 LIL art. 47 Loi respect des principes de la République Directive PNR		
Délibérations CNIL Recommandations CCNE Projets DSA, IA Act		

Une pléthore de termes

Des travaux importants sur la définition du contrôle humain

Intervention humaine (RGPD) … significative (CEDP)	human-in-command (HIC) (idem)
pas sur le “fondement” ou “seul fondement” d’un traitement automatisé (LIL art. 47)	Meaningful Human Control (MHC); contrôle humain, jugement humain (droit international humanitaire)
contrôle humains significatifs, supervision humaine à tout moment (Parl. europ. résol. 20 oct 2020)	moyens non-automatisés (CJUE; Directive PNR; loi respect des principes de la République; projet règlement DSA)
contrôle humain, contrôle effectif (projet de règlement AI Act art. 14)	examen humain significatif (meaningful human review) (Loi de l’Etat de Washington)
human-in-the-loop (HITL) (lignes directrices HLEG)	garantie humaine (recommandations CCNE bioéthique)
human-on-the-loop (HOTL) (idem)	

Quand et comment s'effectue l'intervention?

interventions "système"

Intervention humaine concernant le système dans sa globalité	Intervention humaine concernant une décision individuelle
La définition du problème à résoudre et sa traduction en objectif mathématique	Un décideur humain prend en considération une sortie algorithmique parmi d'autres éléments pour fonder une appréciation globale différente de ce que vise l'algorithme
La définition des critères de décision (pour les systèmes d'IA symbolique)	L'algorithme propose une décision, le décideur humain la valide ou pas en fonction d'autres informations à sa disposition
Le choix des données d'apprentissage	L'algorithme propose une décision, le décideur humain la valide ou pas **sans** consulter d'autres informations
Le nettoyage des données, leur étiquetage (pour l'apprentissage supervisée)	L'algorithme prend une décision avec effet automatique, l'humain surveille le système et l'arrête en cas d'anomalie
Le choix des seuils de sensibilité, de confiance	L'algorithme prend une décision avec effet automatique, l'humain réexamine la décision en cas de contestation
Priorisation de certaines erreurs/biais	
Tests, audits	

interventions "individuelles"

ex ante

ex post

Un contrôle humain pour quoi faire?

1. Détecter et corriger des erreurs
2. Garantir un processus de décision respectueux de valeurs humaines
3. Fixer les responsabilités et démontrer la conformité

Objectif 1: détecter et corriger les erreurs

La CJUE: PUISQUE l'algorithme va générer des faux positifs, ALORS il faut un examen humain de chaque décision

Pas si simple!

- Éliminer tous les biais est impossible; un modèle performant devra s'accommoder de certains biais
- Un certain taux d'erreur est inévitable, même souhaitable
- Dans certains cas, l'intervention humaine sera inefficace, voire pire

Erreurs non-voulues	Erreurs "by design"
Données d'entrée inexactes	Biais statistiques assumés
Biais non-voulus (fausses corrélations; biais de genre…)	Priorisation de faux positifs ou faux négatifs
	Prédictions statistiquement exactes mais fausses dans un cas individuel

Types d'erreurs	Intervention humaine 'système'	Intervention humaine 'décision individuelle'
ERREURS NON-VOULUES		
Erreurs dans les données d'entrée	✅ nettoyage des données	✅ EX POST en cas de contestation
Fausses corrélations	✅ tests en amont, audits en aval	❌
Autres biais statistiques	✅ idem	❌
Prédiction statistiquement exacte, mais contraire à une norme morale ou juridique	✅ imposer des contraintes sur les sorties algorithmiques	✅ EX POST en cas de contestation
ERREURS 'BY DESIGN'		
Cas 1: on ne dispose pas d'une vérité de référence, mais l'humain dispose d'autres informations non-prises en compte par l'algorithme	❌	✅ EX ANTE si l'humain a accès des informations pertinentes non prises en compte par l'algorithme ✅ EX POST en cas de contestation
Cas 2: on ne dispose pas d'une vérité de référence, et on dispose des mêmes informations que l'algorithme	❌	✅ EX ANTE **seulement si** l'humain est aussi performant que l'algorithme (reco. d'image) ✅ EX POST même contrainte

Il ne faut (presque) jamais demander à un humain de faire la même chose que l'algorithme

L'exception à la règle

Les interactions homme-machine

Le problème de l'humain "out of the loop" et de l'inattention

La perte d'acuité et d'intelligence sans entraînement

Les biais de l'automatisation

- confiance excessive
- méfiance excessive

Les pièges de l'explicabilité - anchoring et confirmation bias

Défis de compréhension, et l'effet surprise: "ordinateur tu fais quoi, là?"

La nécessité de décomposer les tâches et définir le rôle de chacun

Ces sujets sont étudiés dans le contexte de l'aéronautique, des voitures autonomes, et des armes autonomes

Objectif 2: Garantir un processus de décision respectueux des valeurs humaines

L'intervention humaine n'est pas seulement un moyen pour atteindre une décision non-erronée, elle a une valeur en soi

Une procédure de qualité contribue à la légitimité et l'acceptabilité de la décision

La personne participe, et comprend

Les valeurs de procédure sont particulièrement importantes lorsque les droits fondamentaux sont en jeu

… et lorsqu'il est difficile de mesurer la qualité de la décision elle-même, notamment en l'absence de vérité de référence (*ground truth*)

Une procédure équitable à géométrie variable ("*some kind of hearing*")

Objectif 3: Fixer les responsabilités et démontrer la conformité

Les programmes de conformité et de responsabilisation mettent l’accent sur les moyens mis en oeuvre

- RGPD
- Lutte anti-blanchiment

L’intervention humaine fera partie des obligations de diligences qui pèseront sur les fournisseurs et utilisateurs d’algorithmes

Absence de contrôle humain = faute

Le contrôle humain sera une “brique” du programme de conformité

Quels leviers pour agir sur la qualité d’une intervention humaine?

L’étalon d’or en termes de qualité	Les leviers pouvant affecter la qualité
La décision collégiale d’un tribunal qui prendrait en considération la recommandation algorithmique en tant qu’ élément de preuve parmi d’autres, avec l’ensemble des garanties d’un procès équitable	Le temps disponible au décideur humain Les autres informations disponibles au décideur La formation du décideur L’indépendance et l’impartialité Une pluralité de décideurs et/ou la confrontation des points de vue (explications de points de vues alternatives) La participation des personnes affectées (procès équitable)

Quelles propositions pour l'IA Act? (1/2)

Le principe	Sa mise en oeuvre
1. Exiger une intervention humaine pour toute décision affectant les droits et libertés individuels	Imposer le principe directement sur l'utilisateur (actuellement les mesures de contrôle humain seront définies par le fournisseur et décrites dans la notice d'utilisation) L'idée d'une intervention humaine **obligatoire mais à géométrie variable** Lever la confusion liée au champ d'application de l'article 22 RGPD
2. Préciser les objectifs de l'intervention humaine	Les modalités de cette intervention seront fixées par le fournisseur et l'utilisateur afin de ● minimiser les risques associés aux décisions erronées ● préserver les valeurs humaines liées à la procédure ● établir une chaîne de responsabilisation pour l'exploitation de l'algorithme

Une méthode inspirée de la régulation des communications électroniques

Préciser les objectifs à atteindre et fournir une boîte à outils + une feuille de route pour les atteindre

Quelles propositions pour l'IA Act? (2/2)

Le principe	Sa mise en oeuvre
3. En fonction du cas d'usage construire les modalités d'une intervention humaine efficace Une intervention humaine à géométrie variable	- une division claire entre les tâches confiées à l'algorithme et celles confiées à l'humain - définir les informations sur lesquelles l'humain s'appuie - définir le timing de l'intervention - préciser la formation des personnes - une explicabilité conçue pour favoriser l'intervention humaine et contrebalancer les biais de l'automatisation
4. Intégrer l'intervention humaine dans l'analyse d'impact et dans la démarche de co-régulation	- Intégrer le choix des modalités d'intervention humaine dans les analyses d'impact effectuées par le fournisseur et l'utilisateur de l'algorithme - Lignes directrices du comité européen de l'IA

Ces travaux feront l'objet d'un papier de recherche publié en mars 2022

Merci!

Winston Maxwell

Telecom Paris, Institut Polytechnique de Paris
winston.maxwell@telecom-paris.fr

telecom-paris.fr/ai-ethics