Law, Society & AI
Research Seminar Series
9 March 2022

Meaningful human review of algorithmic results

Winston Maxwell

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

telecom-paris.fr/ai-ethics

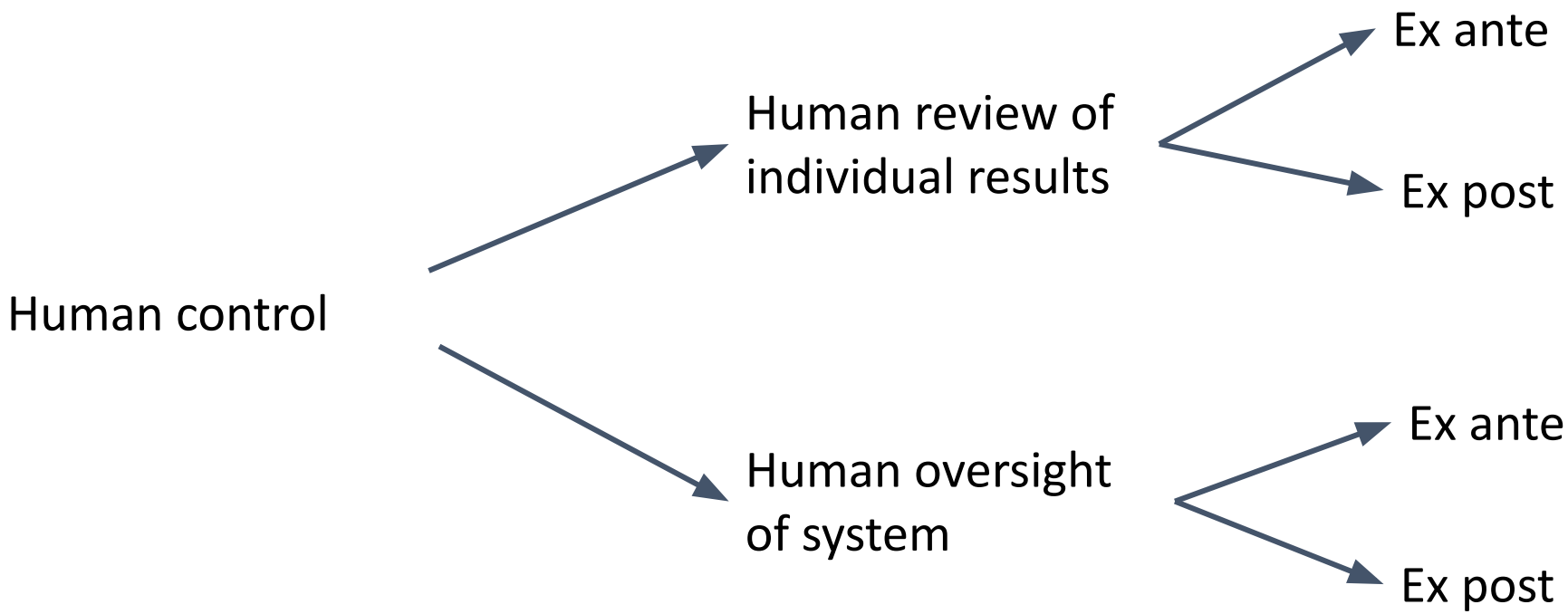
*This presentation is based on an article, currently under peer review, entitled "**Le contrôle humain pour détecter des erreurs algorithmiques**". If you would like a copy, please let me know. Comments welcome!*

This research is conducted in the context of Télécom Paris's research chair "Explainable AI for Anti-Money Laundering" XAIforAML; ANR 20-CHIA-0023

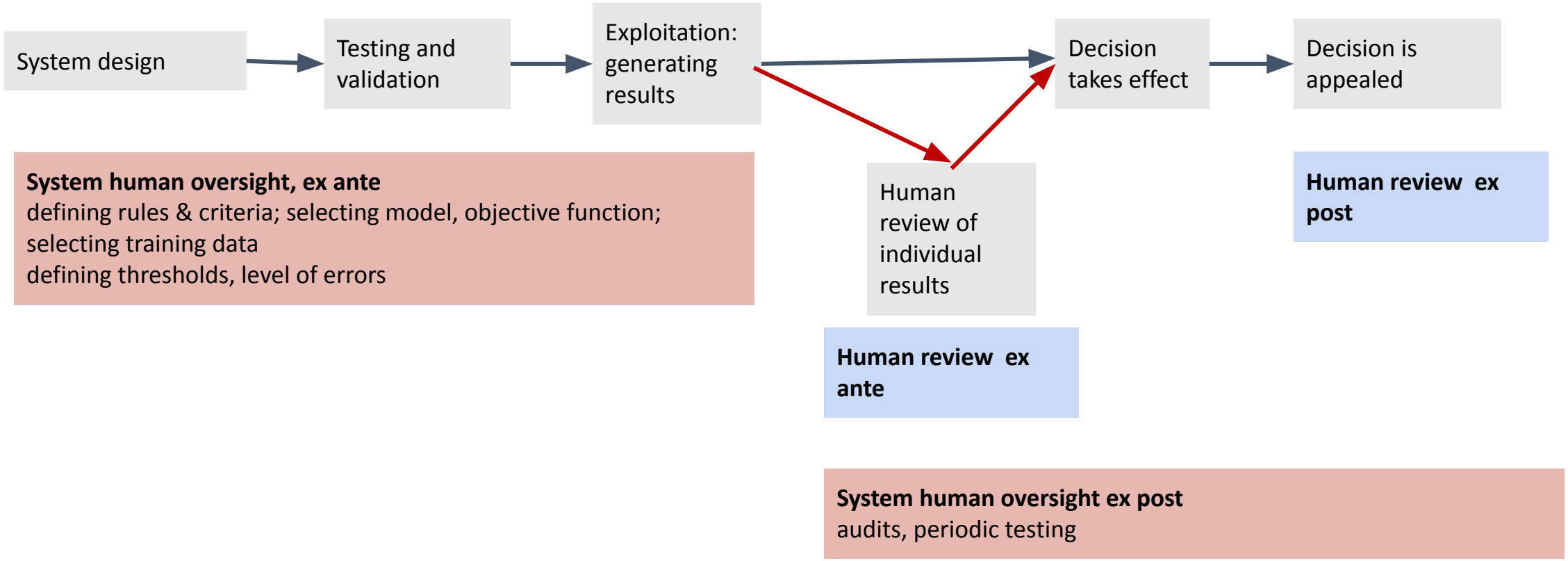
- Definitions
- Human review and oversight in the AI system lifecycle
- Why do we need human review of individual results?
- What are the legal requirements for human review and oversight?
- What algorithmic errors are we worried about, and can individual human review detect them?
- What are the obstacles to effective human review?
- What are the regulatory solutions?

Definitions

Terms used in laws and recommendations
Human intervention
Human oversight
Human review
Meaningful oversight
Meaningful human review
Meaningful human control
Meaningful human review, judgment, intervention and control
Meaningful human intervention and review
Full human oversight at any time
Impartial human review
Under user control
Individual review by non-automated means
Individual re-examination by non-automated means
Effective human oversight
Human oversight and verifications (terrorist content online)
Human-in-the-loop
Human-on-the-loop
Humen-in-command
Decisions…not solely taken on the basis of automated means
Human guarantee

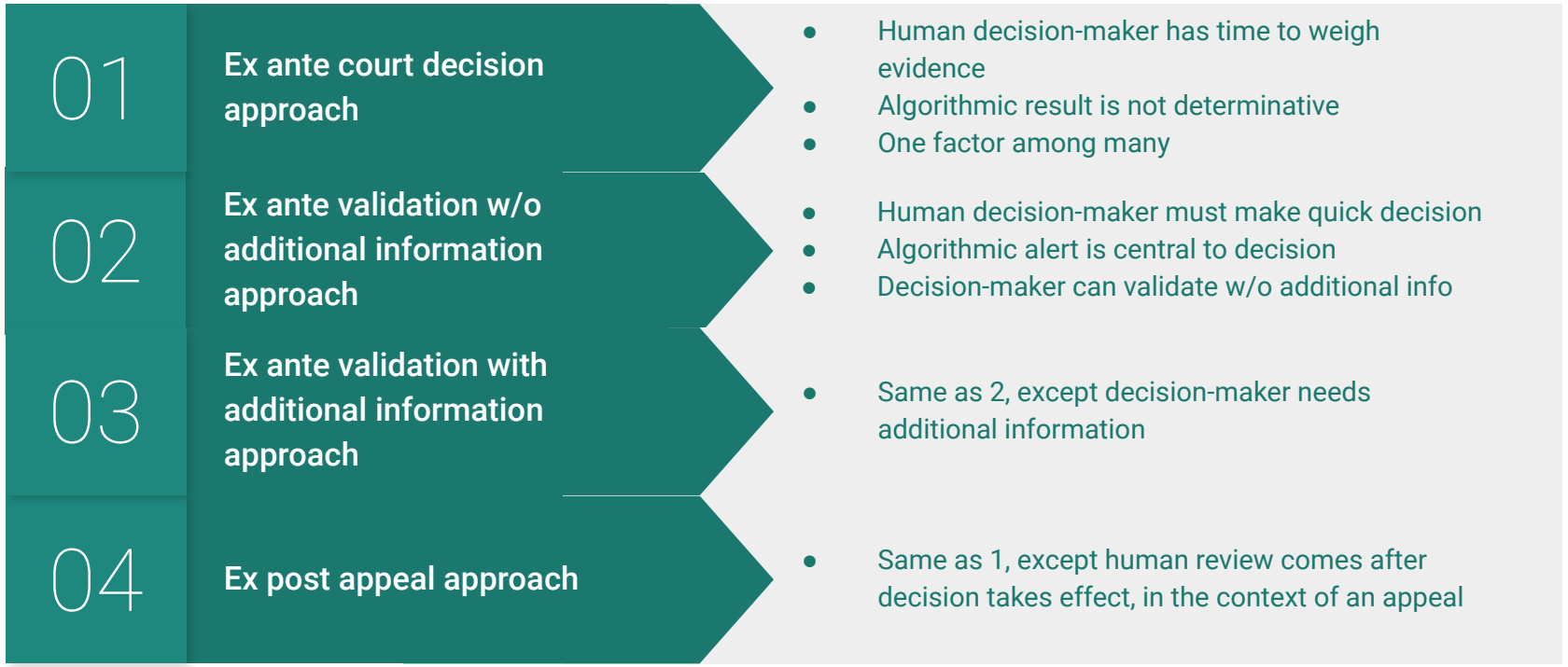

Human control
Human review of individual results
Ex ante
Ex post
Human oversight of system
Ex ante
Ex post

Human review and oversight in the AI system's lifecycle



Four approaches to human review of individual algorithmic results

The most problematic



Why do we need human review?

My paper focuses on human review to detect errors

1. Detect errors, reduce false positives
   - CJEU decisions
   - Facial recognition law
   - Online terrorist content regulation
2. Provide a decision process respectful of human values
   - hearing, due process, presenting arguments, the impression of being heard
   - role reversal
   - GDPR
3. Attribution of moral and legal responsibility for decisions that harm people
   - Literature on LAWS (lethal autonomous weapon systems)
4. Demonstrate compliance, quality control and accountability
   - Standard of care and governance may require human review

What are the legal requirements for human review and oversight?

Europe	United States	Other international treaties
Human review ex ante: CJEU Opinion 1/15 PNR CJEU Case C-511/18 La Quadrature du Net PNR directive 2016/681 Online terrorist content regulation 2021/784 Proposed AI Act (biometric ID only) **Human review ex post:** GDPR Convention 108+ French law on respect of republican principles French law of 6 Jan. 1978 Proposed Digital Services Act **System oversight:** French Constitutional Council decisions CJEU Opinion 1/15 PNR CJEU Case C-511/18 La Quadrature du Net Proposed AI Act	**Human review ex ante:** State of Washington law on facial recognition US Constitution due process clause	International humanitarian law Vienna Convention on Road Transport

Errors in classification algorithms

- Prediction errors in classification algorithms are either false positives or false negatives. True positive rate is a way of measuring predictive performance. FP/FN tradeoffs.
- Human review generally focuses on false positives, not false negatives.
- Problems of class imbalance - base rate fallacy
- Prediction errors can be systematic, in which case we call them bias, or non-systematic, in which case we call them random errors.
- Both bias and random errors cannot be totally eliminated. There is an optimal level, otherwise there's overfitting.

- Types of bias include selection (or representation) bias, historical bias, measurement bias, learning (choice of model) bias, deployment bias.
- Types of random errors may include incorrect data inputs, and variance.
- Some predictions are accurate from a statistical standpoint, but not accurate when applied to an individual case. Is that an "error"?

“All models are wrong, but some models are useful” (G. Box, 1976)

Errors are inevitable, even desirable, in a well-functioning algorithm

The question is how many, and what kind

Is human review effective?

Ex ante human review of algo results without additional information

Human intervention relies on same information as that analyzed by the algorithm.

Effective only for image recognition or short text, where images, or words, speak for themselves.

Example: facial recognition, images of traffic violations

NOTE:

1. will not detect algorithmic biases
2. human automation biases remain a problem

Ex ante human review of algo results with additional information

Human intervention compares algorithmic result with other information to evaluate coherence
Examples: AML alerts, terrorist alerts, fraud alerts

Efficacy depends on quality and accessibility of additional information

NOTE:

1. will not detect algorithmic biases
2. human automation biases remain a problem
3. Class imbalance, base rate fallacy

Except in the case of easy image or text recognition,

the value added of human review of individual results is not obvious.

It needs to be critically assessed.

Obstacles to effective human review

Human bias in face recognition

- performance better for people you know
- performance better for people of your own ethnicity
- performance poorer if rely on algorithmic recommendation ("automation bias")

Human bias in judicial contexts

- judges mete out harsher punishment before lunch
- judges' own emotions and prejudices interfere
- judges systematically follow algo recommendations when they're lenient but will systematically soften harsh recommendations.
- See articles of Cass Sunstein and Aziz Huq

Other biases and cognitive limitations

- alert fatigue (medical diagnosis and treatment)
- algorithmic aversion (medical diagnosis and treatment)
- complacency bias
- human-out-of-the-loop (aeronautic and self driving cars)
- human surprise (aeronautic)
- loss of human skills/reflexes
- human inability to process many variables
- cultural and language bias
- liability regime may bias outcomes

Humans can be bad decision makers, particularly when teamed with computers.

Bad AI + Bad human = Very Bad Outcome!

What regulatory solutions?

Ex post human review will always be useful when individual rights are at stake.

System-level oversight will always be useful, to detect bias, clean data, adjust error thresholds etc.

Ex ante human review is generally useful for recognizing images and short texts from everyday life.

Utility of ex ante human review in other situations is highly dependent on 1) definition of human's and machine's respective tasks, and 2) availability of other information.

To maximize the value of human intervention, the human should be responsible for a task that is different from, but just as important as, the computer's task. (Zerilli et al.)

With certain limited exceptions (e.g. facial recognition), you should never ask a human to do the same thing as the computer, particularly a task that the computer does better!

Except for easy image recognition

never ask the human to do the same task as the computer

What regulatory solutions?

Kind of human review or oversight	Regulatory solutions
Human oversight at a systems level	Current draft of European AI Act seems sufficient
Human review of individual results ex post, in appeals	Key gap is to import certain due process/fair trial principles into private appeal mechanisms ● notice of reasons ● ability to bring new information ● impartial decision maker ● reasoned decision

➢ Due process/fair trial elements are needed in ex post human review (appeals) processes

What regulatory solutions?

Kind of human review or oversight	Regulatory solutions
Human review of individual results ex ante WITHOUT access to additional information	Should be limited to simple image recognition, or short texts
Human review of individual results ex ante WITH access to additional information	Details must be defined in AI system risk assessment: ● separate but equal tasks assigned to computer and human reviewer. Never the same! ● additional information that human reviewer must consult ● decision path for human reviewer when she sees incoherence ● training requirement ● explainability tools to help (but explainability can be a double-edged sword!)

- ➢ Human review processes (and their value added) need to be assessed in AI risk/impact assessment
- ➢ Explainability is double-edged sword

Conclusions

- Most error problems are best treated at the system level
- Systematic human review of algorithmic results before making a decision sounds like a good idea, but it's effectiveness requires:
  - clear separation of tasks between computer and human
  - human evaluates coherence between algorithmic result and other qualitative information
  - decision path and liability of reviewer for actions are defined in advance
- Simple image recognition or text classification may be exception
- Human review in appeals context (ex post) requires some due process/fair trial guarantees

THANK YOU!

Winston Maxwell

Telecom Paris, Institut Polytechnique de Paris

winston.maxwell@telecom-paris.fr

telecom-paris.fr/ai-ethics