Big data – Part 3

Course title – Intitulé du cours	Big Data - Part 3: Behavioral Science for the Ethics of AI
Level / Semester – Niveau /semestre	M2 / S2
Teacher – Enseignant responsable	Jean-François Bonnefon
Other teacher(s) – Autre(s) enseignant(s)	
Other teacher(s) – Autre(s) enseignant(s)	
Other teacher(s) – Autre(s) enseignant(s)	
Other teacher(s) – Autre(s) enseignant(s)	
Other teacher(s) – Autre(s) enseignant(s)	
Lecture Hours – Volume Horaire CM	12
TA Hours – Volume horaire TD	
TP Hours – Volume horaire TP	
Course Language – Langue du cours	English
TA and/or TP Language – Langue des TD et/ou TP	

Teaching staff contacts – Coordonnées de l'équipe pédagogique :

Jean-francois.bonnefon@tse-fr.eu – Office T479 – https://jfbonnefon@github.io

Send questions by email, I try to respond within 48 hours

Course Objectives – Objectifs du cours :

Ethical AI aims at deploying systems which do not create harm, and there are a great many ways they can. This course treats ethical AI as a behavioral problem before it is a technical one. Models learn from human data, then feed outputs back into human behavior. We will look at both directions: how to gather human judgments and use them to inform the alignment of systems; and how apparently helpful systems can still steer ordinary users toward sketchy or harmful actions. We'll dissect notorious cases (e.g. self-driving moral trade-offs), map the kind of behavioral impact that makes a system prohibited under the AI Act, and practice a red/blue design exercise: first invent a maximally unethical yet plausible product, then dismantle someone else's.

The goal is to give you a sharp awareness of practical ethical issues for AI and data science. We will focus on quantitative, measurable outcomes, which means we will often grapple with the problem of translating vague moral concepts into observable metrics.

By end of course you will be able to:

1. Map the areas in which behavioral research and expertise is needed to ensure ethical AI systems
2. Understand the challenges of eliciting human preferences for aligning AI systems
3. Explain pathways by which an AI product could enable or escalate unethical user behavior without malicious intent
4. Identify the behavioral impacts and mechanisms that make an AI system 'prohibited' under the AI act
5. Collaborate in a red/blue exercise of creating a plausible unethical system, then neutralizing another team's system

Prerequisites – Pré requis :

None

Practical information about the sessions – Modalités pratiques de gestion du cours :

Sessions 1 to 4 are lectures. The livelier they are the better: don't be shy about participating! Sessions 5 (red team) and 6 (blue team) are practical exercises where you work in groups under my supervision. You **must** attend these two sessions to get your grade, see below.

Grading system – Modalités d'évaluation :

The red and blue team in-class exercises each count for 50% of your grade.

Bibliography/references – Bibliographie/références :

No prior reading is required, but the lectures will include references you can look into if you want a deeper understanding of the issues we will discuss.

Session planning – Planification des séances

1. Why does ethical AI require behavioral science? Introduction and practical information.
2. Collecting data for ethical alignment. The self-driving car saga.
3. Making people do bad things and doing bad things to people. Unethical impact without malicious intent.
4. What makes an AI system maximally unethical? Prohibited systems under Article 5 of the AI Act.
5. Red team session: Design a very devious but very plausible system.
6. Blue team session: Neutralize or mitigate the system designed by another team in the previous session.

Distance learning – Enseignement à distance :

No distance learning is possible in this class.