

Philosophy & Ethics of Artificial Intelligence

Master in Philosophy of Science

Philippe van Basshuysen

16 December 2019

Worksheet 8 – Ethical Decision-Making with Algorithms

1. **Predictive analytics**

Use the confusion matrix below to explain what type 1 and type 2 errors are.



In what ways might decision-making in predictive analytics be made fairer? Give an example for each possible way.

1. **More Advanced**

Explain in your own words the fairness criteria stated in Chouldecheva (2017).

What is the main problem with these criteria?

1. **Next Week: Designing Ethical Algorithms?**

Readings:

Stuart Russell (2019). *Human Compatible: AI and the Problem of Control*. Allen Lane. Chapter 7: AI: A DIFFERENT APPROACH and Chapter 8: PROVABLY BENEFICIAL AI

Optional readings:

Peter Eckersley (2019). Impossibility and Uncertainty Theorems in AI Value Alignment. *Proceedings of the AAAI Workshop on Artificial Intelligence Safety* 2019

Nick Bostrom (2014). *Superintelligence*. Oxford: Oxford University Press. Chapter 12: ACQUIRING VALUES