Professional Ethics in Computing – Seminar 9

Trustworthy AI

Analysis in Python:

```
Basic frequencies of characteristics:
  p(a1=1) = 0.51
  p(a2=1) = 0.21
  p(a3=1) = 0.1

Test for fairness type 1:
  p(r=1 | a1=0) = 0.32653061224489793
  p(r=1 | a1=1) = 0.3333333333333333
  p(r=1 | a2=0) = 0.21518987341772153
  p(r=1 | a2=1) = 0.7619047619047619
  p(r=1 | a3=0) = 0.3111111111111111
  p(r=1 | a3=1) = 0.5

Test for fairness type 3 in a1:
  p(y=1 | r=0) = 0.5970149253731343
  p(y=1 | r=0, a1=0) = 0.696969696969697
  p(y=1 | r=0, a1=1) = 0.5
  p(y=1 | r=1) = 0.5757575757575758
  p(y=1 | r=1, a1=0) = 0.625
  p(y=1 | r=1, a1=1) = 0.5294117647058824

Test for fairness type 3 in a2:
  p(y=1 | r=0) = 0.5970149253731343
  p(y=1 | r=0, a2=0) = 0.5967741935483871
  p(y=1 | r=0, a2=1) = 0.6
  p(y=1 | r=1) = 0.5757575757575758
  p(y=1 | r=1, a2=0) = 0.5882352941176471
  p(y=1 | r=1, a2=1) = 0.5625

Test for fairness type 3 in a3:
  p(y=1 | r=0) = 0.5970149253731343
  p(y=1 | r=0, a3=0) = 0.5967741935483871
  p(y=1 | r=0, a3=1) = 0.6
  p(y=1 | r=1) = 0.5757575757575758
  p(y=1 | r=1, a3=0) = 0.6071428571428571
  p(y=1 | r=1, a3=1) = 0.4
```

So we find:

- A1 is fair by Independence definition.
- A2 is (approximately) fair by Sufficiency definition, but not Independence.
- No fairness criterion for A3.