Value Alignment NOW!

Prof. Toby Walsh
UNSW Australia | Data61 | TU Berlin

Value alignment

“The most convincing argument [for the problem with an intelligence explosion] has to do with value alignment: You build a system that’s extremely good at optimizing some utility function, but the utility function isn’t quite right.”

Stuart Russell 2014

Value alignment

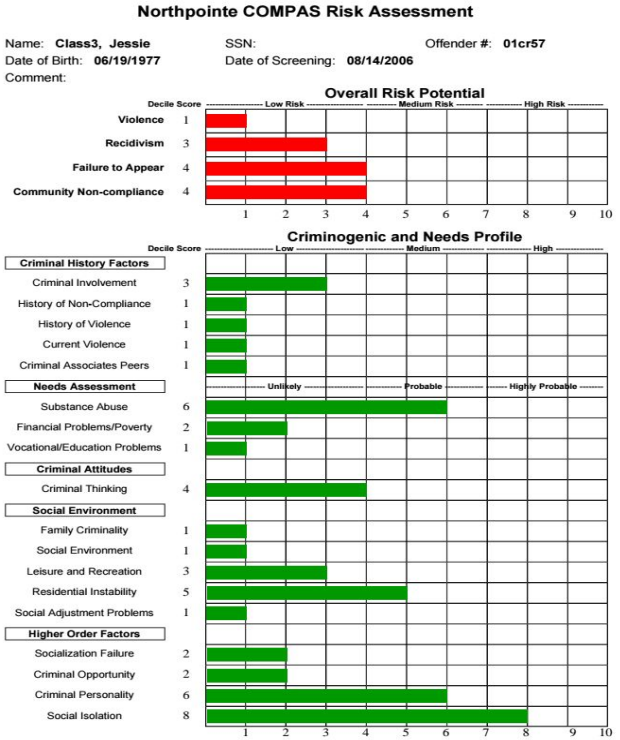
Some experts have expressed concern, though, that it [the invention of superintelligence] might also be the last [event in human history], unless we learn to align the goals of the AI with ours before it becomes superintelligent.

WHY RESEARCH AI SAFETY?, futureoflife.org

Need to align values today

COMPAS misaligned with our values on racial discrimination

Black defendants who did not recidivate over a 2 year period were twice as likely to be misclassified as higher risk compared to white defendants (45 percent vs. 23 percent).



Need to align values today

TAY chatbot misaligned with our values about freedom of speech

At least in Germany, certain aspects of nazism are not tolerated in free speech

Superintelligence may magnify problem

BUT it exists today!

Keep some decisions human

Even if we could correctly predict recidivisim, should we let machines (effectively) lock people up?

Keep some decisions human

Even if we could correctly target, should we let machines kill people?

Need to align values of machines NOW!

Need to align values of machines NOW!

(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

Figure 1. Sample ID photos in our data set.

We need to take responsibility

Some things we can do

That we shouldn’t!