The Incentives that Shape Behaviour

Ryan Carey*, Eric Langlois*, Tom Everitt, Shane Legg
Incentives

Images created by: Anton Barbarov, ironcheese, Graphic Tigers, and Three Six Five
Incentives
Incentives
Incentives

Images created by: Anton Barbarov, ironcheese, Graphic Tigers, and Three Six Five
Incentives

Images created by: Anton Barbarov, ironcheese, Graphic Tigers, and Three Six Five
Incentive Analysis

Safety properties based on the training objectives, not the model architecture
Control Incentives

Age -> Ad Choice

Ad Choice -> Clicks -> Profit

Ad Choice -> Brand Recognition

Browser -> Ad Choice
Response Incentives

- Age
- Browser
- Ad Choice
- Clicks
- Brand Recognition
- Profit