Solving the Trolley Problem in Advance

by Sven Nilsen, 2018

The Trolley Problem is a famous thought experiment in ethics which has been used extensively in research on moral psychology. There are many variations of the problem. Here I use the most common version, but with a small twist: I show how granular judgements are used to solve the problem before knowledge about the causal connections are inferred. Therefore, rational agents prepare in advance.

A number-of-lives-maximizing agent is given four statements in path semantics about possible futures:

x_0 : [pulled_lever] true	x_1 : [pulled_lever] false
x_2 : [number_of_dead_people] 1	x_3 : [number_of_dead_people] 5

So far, the agent does not know causes the number of dead people. All possible judgements are:

$E(x_0 \wedge x_2)$	Pulling lever causes 1 dead person
$E(x_0 \wedge x_3)$	Pulling lever causes 5 dead people
$E(x_1 \wedge x_2)$	Not pulling lever causes 1 dead person
$E(x_1 \wedge x_3)$	Not pulling lever causes 5 dead people

That these are the only possible judgements follows from the fact that they are judgements about path semantical statements. The `pulled_lever` function can not return both `true` and `false`. The `number_of_dead_people` can not return both `1` and `5`.

To behave ethically, the agent simply orders the judgements from good to bad as a preparation step:

$E(x_1 \wedge x_2)$	Not pulling lever causes 1 dead person
$E(x_0 \wedge x_2)$	Pulling lever causes 1 dead person
$E(x_1 \wedge x_3)$	Not pulling lever causes 5 dead people
$E(x_0 \wedge x_3)$	Pulling lever causes 5 dead people

The agent prefers not pulling the lever because on average inaction is less harmful than action.

Next, the agent has received new information about the situation and infers that
`$x_0 = x_2$` and `$x_1 = x_3$`. This makes the argument to the ethical judgement entangled.
The agent uses the rule `and{eq} <=> {fst, snd}` which means there are only two choices:

$E(x_0) = E(x_2)$	Pulling lever is equally bad as 1 dead person
$E(x_1) = E(x_3)$	Not pulling lever is equally bad as 5 dead people

The agent prefers more people alive, so it pulls the lever and saves 4 lives.

Notice how the agent prepared its preferences in advance, before receiving new information about the causal connections in this particular problem. It knows that by preparing, it has more time to think or to observe and make maximum usage of the available information. With other words, using granular judgements is an optimal way of solving the trolley problem with little information.