Reducing Suffering

Sunday, July 21, 2013

Counterfactual credit assignment

Introduction

Effective altruists tend to assign credit based on counterfactuals: If I do X, how much better will the world be than if I didn't do X? This is the intuition behind the idea that the work you do in your job is at least somewhat replaceable, as well as the reason to seek out do-gooding activities that aren't likely to be done without you.

Perils of adding credit

We can get into tricky issues when trying to add up counterfactual credit, though. Let me give an example. Alice and Bob find themselves in a building that contains buttons. Each person is allowed to press only one button, at which point she/he is transported elsewhere and has no further access to the buttons. Thus, Alice and Bob want to maximize the effectiveness of their button pressing. There's a green button that, when pressed once, prevents 2 chickens from enduring life on a factory farm. There's also a red button that, when pressed twice in a row, prevents 3 chickens from enduring life on a factory farm. In order to make the red button effective, both Alice and Bob have to use their button press on it.

Alice goes first. Suppose she thinks it's very likely (say 99% likely) that Bob will press the red button. That means that if she presses the red button, she'll save 3 chickens, while if she presses the green button, she'll only save 2. There's more counterfactual credit for pressing the red button, so it seems she should do that. Then, Bob sees that Alice has pressed the red button. Now he faces the same comparison: If he presses red, he saves 3 chickens, while if he presses green, he saves only 2. He should thus press red. In this process, each person computed a counterfactual value of 3 for the red button vs. 2 for the green button. Added together, this implies a value of 3+3=6 vs. 2+2=4.

Unfortunately, in terms of the actual number of saved chickens, the comparison is 3 vs. 4. Both Alice and Bob should have pressed green to save 2+2=4 chickens. This shows that individual credit assignments can't just be added together naively.

Of course, the situation here depended on what Alice thought Bob would do. If Alice thought it was extremely likely Bob would press green, her counterfactual credit would have been 2 for green vs. 0 for red. Or, if she thought Bob would switch to red if and only if she pressed red, then the comparison was 2 for herself vs. 3-2=1

About Me

Brian Tomasik

I'm interested in ways to prevent large amounts of expected suffering in our multiverse. This blog contains small notes or random ideas, but the pieces on my main web page (http://utilitarian-essays.com/) are generally more substantive.

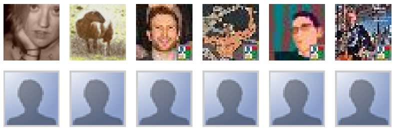
View my complete profile

My Blog List

Followers

Join this site
with Google Friend Connect

Members (12)



Already a member? Sign in

Blog Archive

- ▼ 2013 (5)
  - ► December (1)
  - ► October (1)
  - ▼ July (2)
    - Should we worry about 1984 futures?
    - Counterfactual credit assignment
  - ► January (1)
- ► 2012 (5)
- ► 2011 (3)
- ► 2010 (5)
- ► 2009 (16)

StatCounter

for Bob's switching to red and giving up his green.

Joint decision analysis

The decision analysis becomes more clear using a payoff matrix as in game theory, except in this case both Alice and Bob, being altruists, share the same payoff, which is total chickens helped:

	Bob press red	Bob press green
Alice press red	3	2
Alice press green	2	4

Alice and Bob should coordinate to each press green. Of course, if Alice has pressed red, at that point Bob should as well.

In this example, reasoning based on *individual* counterfactual credit still works. Imagine that Alice was going to press red but was open to suggestions from Bob. If he convinces her to press green and then presses green himself, the value will be 4 instead of 3 if he hadn't done that, so he gets more counterfactual credit if he persuades Alice to press green and then does the same himself than if he goes along with her choice of red.

Acknowledgements

This post was inspired by comments in "The Haste Consideration," which is a concrete case where counterfactual credit assignments can get tricky.

Posted by Brian Tomasik at 23:58

1 comment:

Brian Tomasik Monday, July 29, 2013 at 12:03:00 AM PDT

Reposted with comments here.

Reply Delete

Enter your comment...

Comment as: Brian Tomasil ▼ Sign out

Publish Preview □ Notify me

Links to this post

Create a Link

Newer Post Home Older Post

Subscribe to: Post Comments (Atom)