A computational account of niceness

Ryan Carlson, Michael Brau-Lopez, & Vlad Chituc

4/27/2018

### Models

Here we developed three comptutational models that make judgments of niceness.

* A sacrifice model that judges niceness based purely on 'how much' someone gives, independent of what's given
* A utility model that judges niceness based on how much someone gives of a thing that is valued by the recipient
* A theory of mind model that judges niceness based on beliefs (what an actor thinks the recipient values), desires (what the actor values), and actions (what action the actor takes)

# models can be found in niceness.py on GitHub

### Prosocial scenarios

Here we test these models ability to handle four common prosocial actions:

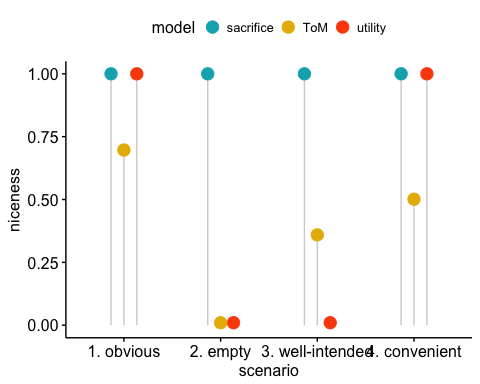
1. An obvious good deed: When an actor gives a gift that both the actor and recipient value Example: when your friend gifts you a really nice scarf that they also like
2. An empty good deed: When an actor gives a gift that they know the recipient does not really value Example: gifting a smelly sock as a joke at a Secret Santa
3. A 'well-intended' good deed : When an actor gives a gift that they believe the recipient values, but in fact does not Example: gifting a winter jacket to someone who is about to move to California
4. A 'convenient' good deed: When an actor gifts something that they do not value themselves, but believe the recipient values. Example: gifting a book when you happen to own two copies

Importantly, while these scenarios focus on gift giving, the themes of these four prosocial scenarios (empty, well-intended, and empty acts of kindness) extend across to a broad variety of prosocial domains such as donating, volunteering, offering emotional support, and cooperating in complex tasks.

# scenarios were tested in niceness.py on GitHub

### Predictions

Below, we visualize the predictions generated from each model of how nice an agent was deemed to be when performing the four prosocial acts above.



### Results

As we can see, all models can handle simple scenarios where a prosocial actor performs an obvious good deed. However, with increasingly complex social scenarios, the Theory of Mind model out-performs simplistic models, both correctly crediting agents' who had good intentions, and penalizing those who gave out of convenience.

### Future directions

* Build a hybrid model that captures some weighted combination of all three models
* Build a model that captures outcomes (both personal benefits for actors and their reactions to those benefits)
* Integrate knowledge of social norms for a given context into the sacrifice model

For more details, questions, or concerns contact [ryan.carlson@yale.edu](mailto:ryan.carlson@yale.edu)