Appendix 1: Instructions

Please read the following instructions carefully. You should not communicate with other participants in the room at any time during the experiment. If you have questions raise your hand and the administrator will assist you.

You will make individual decisions and your payment will depend entirely on your decisions; what other participants do will have no effect on you. In addition to the money you will make in the experiment we will pay you $5 as a show up fee.

All decisions you make are recorded only by the anonymous subject number you received and will only be used for research purposes. Your decisions will remain completely anonymous.

Before describing what the experiment consists of, you are asked to read the information on the following page. After you are done reading it, please move on to the rest of the instructions.

The American Red Cross is empowering people to perform extraordinary acts in the face of emergencies. A hot meal delivered to victims after a disaster, blood when it is needed most, shelter when there is nowhere else to turn, an emergency message delivered to a member of the Armed Forces from their family – these are just some of the ways in which the American Red Cross helps. The American Red Cross is a volunteer-led, humanitarian organization. Each year, it responds immediately to more than 70,000 disasters, including fires, hurricanes, floods, earthquakes, tornadoes. When a disaster threatens or strikes, the Red Cross provides shelter, food, and health services to address basic human needs. In addition, the Red Cross provides assistance to individuals and families affected by disaster to enable them to resume their normal daily activities independently. Here are some examples of how the American Red Cross uses its funds to help disaster victims:

- $25 provides five blankets at an emergency shelter.
- $75 can cover a doctor's visit for an individual injured in a disaster.
- $350 provides emergency food and shelter for 25 disaster victims for one day.
- $2,500 deploys one Emergency Response Vehicle and drivers (including housing and meals for drivers) to a disaster relief operation.

There are two stages of the experiment. The stages are independent and your decisions in one stage will not affect your payoff in the second stage.

For the first stage you will use the three dice and the box you find on your desk. You are asked to throw the three dice simultaneously in the box. Your objective is to get three 5s (5, 5, 5). After each throw you are asked to input in the computer the three numbers you obtained, click "Continue", and wait until you are instructed to roll the dice again. If you input three 5s, you will be prompted to the second stage of the experiment. If you did not input three 5s, the computer will ask you to input the results of the new throw. You can throw the three dice as many as 80 times. Some subjects in the room are asked to throw the three dice 80 times regardless if they obtain three 5s or not.

Your payment could be influenced by how quickly you get the three 5s. You will be rewarded with either $25 or $10. The amount will be randomly determined by the computer and the sooner you obtain three 5s the more likely it is you will be rewarded with $25. The table below describes how your reward will be determined:

If you get three 5s in	Your chance of getting \$25 is:	Your chance of getting \$10 is:
the first throw	90%	10%
the second throw	89%	11%
the third throw	88%	12%
...	...	...
the 32nd throw	59%	41%
...	...	...
the 79th throw	11%	89%
the 80th throw	10%	90%

If you do not get three 5s after 80 throws the computer will reward you the same way as if you would have got the three 5s in the 80th throw (you will have a 10% chance of getting \$25 and a 90% chance of getting \$10).

After you either input three 5s or you reach the limit of 80 throws, you will be asked to follow the instructions on the screen corresponding to the second stage.

For Subjects in Control:

If you finish the experiment before other subjects, you are asked to remain at your desk and to not disrupt the other participants. After the program terminates you will be able to use the computer in any way you like, for example to browse the internet.

For Subjects in Treatment:

In the second stage of the experiment we give you the opportunity to donate a part of your earnings in the first stage of the experiment to the American Red Cross. You can choose to donate none or all of your earnings, as well as any integer amount in between.

The computer will display your total compensation for the experiment: your earnings in the first stage + \$5 show up fee - your charitable donation to Red Cross. When everybody finishes the experiment you will be paid this amount by the experimenter. If you finish before other subjects, you are asked to remain at your desk and to not disrupt the other participants. After the program terminates you will be able to use the computer in any way you like, for example to browse the internet.

After the conclusion of the experiment, everybody's donations will be totaled and a check will be mailed to the American Red Cross. One of you will be randomly selected to walk with the experimenter to the nearest mail box to witness the mailing of the check. For this task, the monitor will be paid an additional 5 dollars.

Appendix 2: Point Estimates and Confidence Intervals for The Effect of Treatment on Cheaters

A subject who never reports rolling a 5-5-5 is certainly honest. A subject who reports rolling three fives is either honest and lucky or dishonest. We can therefore identify a subset of honest subjects, but we cannot identify which of the successful subjects are dishonest. Instead, we perform some calculations and Monte Carlo simulations to estimate how giving in Treatment compares to giving in Control among dishonest subjects.

We decompose the difference in giving among successful subjects observed in the data $\Delta d_{S=1} = E[d|T = 1, S = 1] - E[d|T = 0, S = 1]$ as follows:

$$\begin{aligned}\Delta d_{S=1} &= \text{Prob}(c = 1|T = 1, S = 1) \cdot \{E[d|T = 1, c = 1] - E[d|T = 0, c = 1]\}\\ &+ \text{Prob}(c = 0|T = 1, S = 1) \cdot \{E[d|T = 1, c = 0] - E[d|T = 0, c = 0]\}\\ &+ \{\text{Prob}(c = 1|T = 1, S = 1) - \text{Prob}(c = 1|T = 0, S = 1)\} \cdot \{E[d|T = 0, c = 1] - E[d|T = 0, c = 0]\},\end{aligned}$$

where $\text{Prob}(c = 1|T = 1, S = 1)$ is the probability a subject is a cheater conditional on being in Treatment and reporting a success, and $\text{Prob}(c = 1|T = 0, S = 1)$ is the probability a subject is a cheater conditional on being successful in the control group.

We also observe the difference in giving between successful and unsuccessful subjects in Control. This can be written as: $E[d|T = 0, S = 1] - E[d|T = 0, S = 0] = \text{Prob}(c = 1|T = 0, S = 1) \cdot \{E[d|T = 0, c = 1] - E[d|T = 0, c = 0]\}$. Plugging the resulting expression for $\{E[d|T = 0, c = 1] - E[d|T = 0, c = 0]\}$ back in the decomposition for $\Delta d_{S=1}$ above gives:

$$\begin{aligned}\Delta d_{S=1} &= \text{Prob}(c = 1|T = 1, S = 1) \cdot \{E[d|T = 1, c = 1] - E[d|T = 0, c = 1]\}\\ &+ \text{Prob}(c = 0|T = 1, S = 1) \cdot \{E[d|T = 1, c = 0] - E[d|T = 0, c = 0]\}\\ &+ \frac{\text{Prob}(c = 1|T = 1, S = 1) - \text{Prob}(c = 1|T = 0, S = 1)}{\text{Prob}(c = 1|T = 0, S = 1)} \cdot \{E[d|T = 0, S = 1] - E[d|T = 0, S = 0]\}.\end{aligned}$$

We then solve for $E[d|T = 1, c = 1] - E[d|T = 0, c = 1]$ using point estimates for all the other differences above from Table 2. For example, for donations out of 10 dollars, we observe: $\Delta d_{S=1} = -2.24$, $E[d|T = 1, c = 0] - E[d|T = 0, c = 0] = -1.56$, and $E[d|T = 0, S = 1] - E[d|T = 0, S = 0] = -0.54$. Point estimate for conditional probabilities could be obtained in the following way. 32 out of 64 subjects did not report a success in Control and are therefore certainly honest. The proportion of honest subjects who would fail to report a success after 80 dice rolls is about 73.2%. Thus the expected number of honest subjects in Control is $\frac{32}{0.732} = 43.71$. Of these, $43.71 - 32 = 11.71$ were successful, so $\text{Prob}(c = 0|T = 0, S = 1) = \frac{11.71}{32}$. Similarly, 25 subjects fail to report a success in Treatment, so there must be around $\frac{25}{0.732} = 34.15$ honest subjects in Treatment. Thus, out of the 41 subjects who reported a success in Treatment, only $34.15 - 25 = 9.15$, are honest. Thus, $\text{Prob}(c = 0|T = 1, S = 1)$ is approximately $\frac{9.15}{41}$.

With these values, the equation above for donations out of 10 dollars becomes:

$$-2.24 = \frac{31.85}{41} \cdot \{E[d|T = 1, c = 1] - E[d|T = 0, c = 1]\} + \frac{9.15}{41} \cdot (-1.56) + \frac{\frac{31.85}{41} - \frac{20.29}{32}}{\frac{20.29}{32}} \cdot (-0.54),$$

which yields the point estimate for $E[d|T = 1, c = 1] - E[d|T = 0, c = 1]$ of -2.27. Similarly, the point estimate for $E[d|T = 1, c = 1] - E[d|T = 0, c = 1]$ out of 25 dollars is -3.48.

To obtain confidence intervals for these point estimates we run Monte Carlo simulations. We draw a value for each difference in giving we used above from a normal distribution centered at the point estimate and with the standard deviation estimated from regressions. For conditional probabilities we draw a value from the simulated posterior distribution of the number of cheaters in each group, conditional on observing 32 and 41 subjects reporting a success, respectively. We then solve for the estimate for $E[d|T = 1, c = 1] - E[d|T = 0, c = 1]$ according to the equation above. With 50,000 draws, $E[d|T = 1, c = 1] - E[d|T = 0, c = 1]$ was above 0 around 6% of the time for donations out of 10 dollars, and around 13% of time for donations out of 25 dollars.

Appendix 3: A More General Framework

We examine how sensitive the predictions derived under moral cleansing, 2b, 3b, and 4b, are to the strong assumptions we used in the stylized model we described in the text. We develop a fairly general framework of moral behavior and we show that under reasonable assumptions all predictions derived under moral cleansing hold or are strengthened. We assume that the subjects' ethical choices c and d are correlated to some unobserved underlying propensity towards pro-social behavior, θ. We observe that subjects in Treatment are less honest on average than the subjects in Control. We split the domain of θ in three groups: the subjects who never cheat, AH; the marginal cheaters (or honest subjects), MC; and the subjects who always cheat, AC. We can safely assume that $E[\theta|\theta \in AC] < E[\theta|\theta \in MC] < E[\theta|\theta \in AH]$.

Suppose moral cleansing is the mechanism driving the increase in cheating in Treatment. Then a subject with a given θ should donate more when she cheated than when she did not. We assume throughout that if a subject chooses the same c in both treatments, she will choose the same d in both treatments; that is, the effect of Treatment on giving operates exclusively via its effect on cheating. We can then write the difference in giving between treatments as: $\Delta d = E[d|T=1] - E[d|T=0] = \text{Prob}(\theta \in MC) \cdot \{E[d|T=1, \theta \in MC] - E[d|T=0, \theta \in MC]\} = \text{Prob}(\theta \in MC) \cdot \{E[d|c=1, \theta \in MC] - E[d|c=0, \theta \in MC]\}$. The last expression is positive under moral cleansing, so giving should increase in Treatment under moral cleansing. This is our Prediction 2b, which is rejected by Result 2.

The difference in average giving of honest subjects between treatments can be written as $\Delta d_{c=0} = E[d|T = 1, c=0] - E[d|T=0, c=0] = \text{Prob}(\theta \in MC) \cdot \{E[d|T=1, \theta \in AH] - E[d|T=0, \theta \in MC]\}$. The term in parentheses can be rewritten as $E[d|c=0, \theta \in AH] - E[d|c=0, \theta \in MC]$. Assuming that for the same choice of c, donations d are non-decreasing in θ, the difference is positive because the marginal cheaters have smaller θ on average than the always honest subjects. So donations of honest subjects should increase in Treatment under moral cleansing. This is a strengthening of our Prediction 3b; both are rejected by Result 3.

The difference in average giving of dishonest subjects between treatments can be similarly written as $\Delta d_{c=1} = E[d|T=1, c=1] - E[d|T=0, c=1] = \text{Prob}(\theta \in MC) \cdot \{E[d|T=1, \theta \in MC] - E[d|T=0, \theta \in AC]\}$. The term in parentheses can be rewritten as $E[d|c=1, \theta \in MC] - E[d|c=1, \theta \in AC]$. Like before, the difference is positive because for the same ethical choice in the first stage, the marginal cheaters have a larger average θ than the subjects who always cheat. Donations of dishonest subjects should therefore increase in Treatment under moral cleansing. This is our Prediction 4b, which is rejected by Result 4.