

A note continuing the generalization of the free-will argument
to varying degrees of control over belief.

Assume, for purposes of discussion, that Jane has an “inner self” that tries to change her beliefs about free will. Let Θ be the actual probability that Jane’s inner self will succeed in changing her beliefs when it tries to do so. $\Theta = 0$ is taken to be equivalent, for this purpose, to the nonexistence of free will over what beliefs Jane chooses to hold about free will. Note that the absence of free will in general implies $\Theta = 0$, though $\Theta = 0$ needn’t imply the absence of free will in general, only the absence of free will in the area of choosing beliefs about free will.

Let θ be a variable ranging over possible values of Θ . Jane has some prior subjective probability density function (pdf) $p(\theta)$ on $[0, 1]$ for what she thinks Θ is.¹ Jane’s inner self tries to make Jane hold some other pdf, $t(\theta)$. If her inner self fails, Jane will instead be determined to hold some other pdf, $q(\theta)$.

Following the discussion in the main text, we can consider π_t , the expected value of the probability density that Jane assigns to the correct value of Θ :

$$\pi_t := \int_0^1 p(\theta) [\theta t(\theta) + (1 - \theta)q(\theta)] d\theta.$$

Jane’s inner self wants to choose a pdf t that will maximize π_t . Clearly, it suffices to maximize

$$\int_0^1 p(\theta)\theta t(\theta)d\theta.$$

For instance, suppose $p(\theta) = 1 \forall \theta \in [0, 1]$. Then Jane wants to maximize

$$\int_0^1 \theta t(\theta)d\theta.$$

She can make this quantity bigger by putting lots of the weight of $t(\theta)$ toward high values of θ .

As another example, suppose

$$p(\theta) = \begin{cases} 0 & \text{if } \theta \in [0, \frac{1}{e}) \\ \frac{1}{\theta} & \text{if } \theta \in [\frac{1}{e}, 1]. \end{cases}$$

Now it suffices for Jane’s inner self to maximize

$$\int_{\frac{1}{e}}^1 t(\theta)d\theta.$$

This time, it doesn’t matter what shape t has, so long as all of its density is in the region $[\frac{1}{e}, 1]$.

Perhaps Jane is not so concerned with the probability density she assigns to the correct value but rather with making sure her distance from the correct value isn’t too large. Suppose Jane decides on some error metric $m(\theta, \Theta)$, which measures how far Jane would be from correct by believing θ

¹If we want to allow Jane to assign nonzero probability to $\Theta = 0$, $p(\theta)$ could include a partial Dirac delta function at $\theta = 0$.

when Θ is true. For instance, Jane could take $m(\theta, \Theta) = |\theta - \Theta|$ or $m(\theta, \Theta) = (\theta - \Theta)^2$. For a particular value of Θ , Jane's error in holding some pdf f may be defined as

$$\int_0^1 f(\theta)m(\theta, \Theta)d\theta.$$

So, for some particular Θ , Jane's expected error is

$$\Theta \int_0^1 t(\theta)m(\theta, \Theta)d\theta + (1 - \Theta) \int_0^1 q(\theta)m(\theta, \Theta)d\theta.$$

Since Jane's inner self doesn't actually know what Θ is, it has to consider an expected value over θ , which I've called E_t :

$$E_t := \int_0^1 p(\theta) \left[\theta \int_0^1 t(x)m(x, \theta)dx + (1 - \theta) \int_0^1 q(y)m(y, \theta)dy \right] d\theta.$$

Of course, it suffices to minimize

$$\int_0^1 p(\theta)\theta \left[\int_0^1 t(x)m(x, \theta)dx \right] d\theta.$$

If Jane is concerned with the prudential value of holding various densities, she could take $m(\theta, \Theta)$ not to be a direct measure of error but a measure of the cost of believing θ when Θ is in fact true.