Journal of Economic Literature 2018, 56(1), 1–17
https://doi.org/10.1257/jel.20161260

Morality, Policy, and the Brain†

Aldo Rustichini*

The book Moral Tribes: Emotion, Reason, and the Gap between Us and Them, *by Joshua Greene, invites the reader to give a new look at the foundation of ethics and, by implication, to policy. Its specific strength is the systematic integration of new methods from neuroscience into a very old debate. Having something new and substantial to add in an investigation that has been at the center of the philosophical debate in Western civilization for twenty-five centuries is remarkable. While I invite everyone to read and enjoy this wonderful book, I take here the opportunity to invite economists to take the challenge. We are particularly interested in the question, "Is there a specific contribution that economics can give to this debate?" I believe there is and this insight is now in danger of being lost. This is my attempt to indicate where the research should look now. Maybe it is not too late.* (*JEL* D12, D63, D87, Z13)

1. *Moral Dilemmas*

1.1 *The Story of a Father and a Son*

In *Most* ("The Bridge", a short thirty-three minute 2003 Czech film directed by Bobby Garabedian[1]) the story is told of a father and a son. The father is a humble worker, tender of a railway bridge over a river; his task is raising the bridge when a boat needs to transit in the river, and lowering it when a train is approaching. The father loves the young eight-year-old son, Lada. The initial scenes display the two enjoying their company, detailing the ravishing joy of every moment they spend together. There is no mother in sight, and no explanation, even indirect, is provided for such significant absence. The limelight is on the father and son.

One day Lada says that he wants to spend a day on the bridge with the father, who is reluctant but won over by the witty answers of the son.

> Father: It is going to be dark on the bridge.
>
> Son: We will bring flashlights.
>
> Father: It is going to be cold on the bridge.
>
> Son: We will bring hot chocolate.

And so they go. In the following scenes, the father is tending his job, the son is fishing in the river. A boat passes, and the bridge is raised. The next train is only due to arrive in an hour, and there seems to be no need to rush.

Suddenly the train is shown approaching the bridge, full of travelers. We see their

*Department of Economics, University of Minnesota. Supported in part by grants from the National Science Foundation (SES 1357877). I thank Angelo Rustichini for a very useful insight.

†Go to https://doi.org/10.1257/jel.20161260 to visit the article page and view author disclosure statement(s).

[1]The curious reader can watch it on YouTube, searching for *Most The Bridge*.

faces, all taken in their thoughts, their occupations. Among them, a young woman, a drug addict, is shown getting ready to inject drugs into her veins, with a lost, numbed look on her face. We see the father, distracted and unaware of the coming disaster. We see the son realizing that the train is coming too soon. After trying to attract the attention of the father, he rushes to the bridge, trying to pull a lever that would lower the bridge. He fails and during the attempt falls into the bridge gear works. If the gears are activated, he will be crushed. Only now the father sees the son fall and the train approaching. He is now facing a horrific moral dilemma, and has a few seconds to decide.

2. *Moral Tribes*

The book by Joshua Greene, *Moral Tribes*, begins as a fascinating study of dilemmas like this, how human beings process these dilemmas, and proceeds to show us what lessons in ethics, and in social policy, one can draw from their solutions.

The book has three main theses. The first thesis is that in ethical choices human beings are moved by two systems, one automatic, impulsive, perhaps more primitive, and one more reflexive, computational. The idea is best illustrated by an analogy that the author uses assiduously with dual-mode photographic cameras (see, for example, page 133). The first mode is automatic settings, to be used for typical situations: in this case all the adjustments are made automatically by the camera. The second mode is manual mode: all the settings have to be decided by hand, deliberately, slowly, with a careful consideration of all the possible trade-offs. Similarly, moral thinking has two modes: the "automatic" is impulsive, while the "manual" makes careful comparisons of the good and bad that would follow our choice.

The second thesis is that evolution has endowed us with moral codes that help solve the fundamental social problem of cooperation among otherwise selfish individuals. "Morality evolved as a solution to the problem of cooperation, as a way of averting the tragedy of commons. Morality is a set of psychological adaptations that allows otherwise selfish individuals to reap the benefit of cooperation. . . . The essence of morality is altruism, unselfishness, a willingness to pay a personal cost to benefit others" (Greene 2013, p. 23).

The third thesis is that these codes are very effective to solve the dilemmas within groups (us) but very poorly developed to solve the dilemmas of cooperation among groups (us versus them). The author concludes with presenting his own solution, an elaboration of utilitarianism, that he calls deep pragmatism.

The heart of the book is in the systematic analysis, with completely new methods, of several moral dilemmas. The emphasis on this section in the book is perhaps less than we could expect: but still we call this the heart because it is the place where a new approach is tested. The dilemma artistically presented in the *Most* movie is an example of these dilemmas. We will, for our purposes in this review, call it the *Most* dilemma. *Most* has, naturally, the emotional load derived from the identification of the spectator with the father, and with an interesting additional feature, relatively new in the trolley literature we will review later, that the decision maker is related to one of the potential victims.

3. *Hypothetical Choices*

To understand which new insights Greene and his research have brought to the philosophical discussions, we need to understand how trains and trolleys came to be the fruit fly of ethical reasoning. This survey is necessary to marshal all the alternative ethical theories that have been suggested in this debate. This will be essential in our evaluation later.

We begin the survey with the conclusion of the story of the father and the son. After moments of excruciating doubts, the father decides to pull the lever lowering the bridge. The train passes safely, and the camera shows the passengers totally unaware of the drama, with a single exception: the young woman who sees the distraught face of the father as he runs to assist the son. Perhaps she, somehow, has understood the sacrifice that the father made. There is no doubt about the fate of the son: he is shown, lifeless, in the arms of the grieving father. The moral dilemma has been solved.

Was it the right choice? The movie has an answer. We follow the father, who moves to a different city ("I will discover something. New city. New job. New dreams. New people. New life."). As he wanders in the new city, he sees the young woman on the train, now with a bright look on her face. She is holding a young child in her arms, and talks to him, the image of a happy mother. The father sees them, and smiles. He throws his arms to the sky, exultant and grateful. The camera follows his gaze to the heavens, and the movie ends. A new life has taken the place of the lost son. This is a new religion, with new commandments. The story parallels the sacrifice of the son Isaac that the Lord asks from Abraham (Genesis 22); only the conclusion is very different.

3.1 *Trolley Problems*

The use of hypothetical trolley problems begins with an article by Philippa Foot (1967), in her reconsideration of the moral basis of abortion.

3.1.1 *Double Effect*

Foot begins by inviting readers to an unprejudiced reconsideration of the "doctrine of double effect," proposed by Thomas Aquinas (1265–73, II-IIae, Question 64, "Murder") to address the question of "whether it is lawful to kill a man in self-defense" (article 7). To answer, he proposes, we should first note that actions may have two effects—"one which is intended, the other which is beyond intention." Now, the moral nature of an act depends (*recipiunt speciem*)[2] "on what is intended, not on what is besides the intention." In the case of murder for self-defense, one's intention is to save one's life, slaying the aggressor is accidental. Hence "it is not necessary for the salvation that a man omits the acts of killing in self-defense, since one is bound to take care more of one's life than of another's." One may wonder whether St. Thomas has only shifted the burden of the proof to the assumption that the moral nature of actions depends on intentions, and we think he did. He has a good reason to do this, and we will return to this later.

If one accepts his premise, however, the parallel with abortion that Foot herself draws may provide an answer. The doctrine highlights a difference between an operation of hysterectomy which "involves the death of fetus as the foreseen but not strictly intended consequence," while abortions "kill the child and count as the direct intention of taking an innocent life." The doctrine of the double effect may sometimes lead to unpalatable conclusions. On the other hand, opposing the doctrine may lead to views that are impossible to defend. Foot suggests two dilemmas. In the first, the *magistrate and the riots* problem, a magistrate who faces possible riots that might lead to lynching of five accused, has the opportunity of framing one innocent to save the lives of others. In the second, the *trolley driver* problem, the driver of a runaway trolley can only steer from one track to another; five men are working on one, and a single worker on the other. It seems, and experimental tests confirm the intuition, that one would approve the sacrifice of the single

[2] "Morales autem actus recipiunt speciem secundum id quod intenditur, non autem ab eo quod est praeter intentionem, cum sit per accidens, ut ex supradictis patet."

man in the trolley driver problem, but would instead object to the very same sacrifice in the case of the magistrate. Why? The doctrine of the double effects offers a way out of the difficulty, because in the first case the death of the innocent framed man is a side effect; in the second case the death of the single railroad worker is part of the plan.

3.1.2 *A Stricter Duty*

Foot goes on to examine limitations of the doctrine—and suggests her own, different doctrine, based on the distinction between *the duty to avoid injury* and *the duty of bringing aid*. Her criterion is simple: "To refrain from inflicting injury ourselves is a stricter duty that to prevent other people from inflicting injury" (Foot 1967, p. 4). To illustrate Foot's distinction, Judith Thomson (1985) compares the trolley driver with the *transplant* problem: five patients in a hospital need five different organs for a transplant that will save their lives, and all these organs could be harvested from a single healthy patient who happens to be at the same hospital for a routine checkup. In this case, most people object to the sacrifice of one to save five. Foot's solution of the seemingly contradictory answers to the trolley and the transplant is that in the transplant you choose between killing one and letting five die (and the first is worse than the second, because of the stricter duty clause), whereas in the trolley driver you choose between killing one and killing five (and now the second is obviously worse).

3.1.3 *Killing or Letting Die*

To test this explanation, Thomson proposes a modification of the trolley driver, the *bystander at the switch*. A bystander is observing a runaway unmanned trolley, which is about to run over five people, but can operate a switch diverting it to another track, where it would only kill one. In this choice, operating the switch seems the right choice; but we are clearly preferring killing one to letting five die, which is the opposite of what Foot was assuming.

The difference from the driver dilemma is that here, operating a switch may not seem like killing, as it is when instead you have to steer a trolley toward that person. The difference is even starker in the *footbridge,* which Thomson calls the fat man dilemma. In this case, a trolley running to kill five can only be stopped by shoving a fat man standing on a footbridge onto the tracks. This would stop the trolley (you are too lightweight for this purpose). There is a strong, seemingly natural resistance to save the five at the price of killing the fat man.

3.1.4 *Categorical Imperative*

The footbridge is paradoxical because the balance of lives saved is the same as in the switch, but the answer is the opposite. Thomson dispenses with an easy explanation of this answer in the footbridge dilemma. It is a rather mechanical interpretation of the *categorical imperative* in the second formulation (Kant 1785)[3] with a slight modification of the switch problem called the *loop* problem. In the loop problem a trolley running to kill five people can be diverted by operating a switch to a different track, where it would kill a single person (so far the dilemma is identical to the switch) but this second track then loops back into the other and if unstopped would still kill the five. In this case, the single person killed is used as a mean, contrary to the categorical imperative, but still the right answer seems to use the switch.

So what is really the difference between the switch on the one hand, and the footbridge

[3] "Act in such a way that you treat humanity, whether in your own person or in the person of any other, never merely as a means to an end, but always at the same time as an end." This is the version of the categorical imperative Greene uses in his book, on page 115. The first formulation of the imperative is absent in *Moral Tribes*. As we shall see later, it should not be.

or transplant on the other? Thomson suggests her own criterion: rights.

3.1.5 *Rights Trump Utilities*

In the footbridge problem, the action "push" violates the right to life of the fat man; and "other things being equal, to kill a man is to infringe his right to life, and we are therefore morally barred from killing" (Thomson 1985, p. 1395). In the switch problem, neither of the two choices available violate anyone's rights. Throwing the switch simply changes an event which is threatening five to an event threatening one.

4. *A Summary So Far*

We have listed in a narrow space many dilemmas, and just as many explanations. Some considerations may be necessary at this point.

4.1 *Ambiguity*

First of all, ambiguity is the substance of the dilemmas. In the comment to Philippa Foot's (1967) article, G. E. M. Anscombe reflects on the *drug in short supply dilemma*, a milder version of the transplant, where one has to decide whether to administer the scarce drug to a single patient that was the former intended recipient, or to five newcomers. But why is the effect of the drug so different in the two cases? The story does not tell, and we are left free to speculate. Anscombe (1967) notes that "the latter case is vaguely sketched, and one pictures resources being lavishly used beyond necessity on one" as compared to the effective use for five. This same ambiguity is evident from the discussion of the problems, where in several instances the person who interprets the results notes that the interpretation changes depending on the unobservable completion of the partial description of the environment provided by the investigator (see, for example, the "addition" to transplant, in which we are informed that the ailments of the five patients were induced by malpractice of the surgeon (Thomson 1985, p. 1399)).

The ambiguity induces subjects answering a dilemma to fill in the details, just like the professional philosophers do when discussing the alternatives. The problem is that we do not know which details they add. In terms of the best practice in experimental design, the experimenter in the trolley problems is constantly in danger of, and quite frequently is, losing the control over the experiment. He cannot be sure what the behavior of the subjects means, because he cannot know what the subject thought.

4.2 *Veil of Ignorance*

All decision makers in our hypothetical choices have no stakes in the game. They are all, for practical purposes, behind a veil of ignorance, in the simple sense that they are not directly affected by the outcome, no matter what that is. This is of course the fundamental difference between the father in the *Most* and the deciders in all the dilemmas considered here. They are behind the veil also in the deeper sense that they realize that there is another way of looking at the dilemma: they are setting a precedent with their decision. So they realize that if later they were to find themselves as one of the five persons on the track or five patients needing transplants, or the single healthy patient chancing to be in the hospital where his organs are needed, the hypothetical choice they just made might be relevant for the ensuing situation.

That subjects actually do this can be illustrated by looking at recent applications of these investigations to the problem under current discussion of the *ethical cars*. As cars that drive themselves are becoming a reality, a problem that is discussed is what the program should do in some critical situations. You are in a car driven by a computer program, and the car is in risk of

running over five people on the way. The choice for the computer program is whether to swerve suddenly to the side, where the car will hit a wall and kill you, or to run over the five people. What program would you like? Experimental tests have been run (Bonnefon, Shariff, and Rahwan 2015) following the protocol described for the trolley problems. They show that people approve programs that take the utilitarian calculation in mind and spare the five, and say that they favor a wide use of these programs. But they are also unlikely to *buy* those programs, so they are (consciously or unconsciously) perfectly aware of the implications of the legal provisions and their choice as consumer, and the two choices differ.

5. *What Brain Scans Teach Us*

Of all the complexity of these problems, Greene and coauthors focus on the contrast between the switch and the footbridge.

5.1 *Hypotheses*

In Greene et al. (2001) (but see also Greene and Haidt 2002 and Greene et al. 2004), the authors contrast the loop problem and the footbridge (so that there is no "man used as mean" confound). They suggest that the explanation of the difference is the emotional saliency of the choice in footbridge, where you have to kill a man with your own hands, rather than through the intermediation of the switch. A more general hypothesis is that some moral decisions involve emotional processes more than others, and that this explains the inconsistencies we detect. This is close to the stricter duty of Foot, but with an additional and substantial difference that the command to refrain from inflicting injury ourselves is shrouded with emotional valence.

5.2 *Predictions*

The authors propose an experimental test, formulating several predictions that should hold if the hypothesis is true. First, brain areas associated with emotions should be more active when subjects are contemplating choices like the one in footbridge. Second, in those subjects who eventually choose "push" in footbridge, the choice is reached after the resolution of a conflict in which an automatic response ("Thou shalt not kill") is overridden by the calculation that choosing otherwise will result in five live people and only one dead. This should induce patterns of brain activation similar to those observed in the incongruous Stroop task. In the Stroop task, a subject is required to state, for example, the written color of the word "green" written in red, where stating the color is the automatic response and reading the word is the deliberate, reflexive choice. Finally, in the trials where the choice is incongruent with the emotional response (for example, declaring "push" appropriate in footbridge) the decision should be reached after a longer deliberation necessary for formulating the automatic response, countering it and choosing the opposite; thus, longer response times.

5.3 *Design and Results*

The experimental design asks subjects to consider a suggested action and declare it appropriate or not. The choices are either nonmoral (for example, "You can travel by train or bus," with suggested action "bus"); moral non-personal (choices with moral implications, with no personal direct involvement in the action, such as the switch or the loop dilemma); and finally moral and personal (the footbridge or the transplant dilemma). The papers Greene et al. 2001 and Greene et al. 2004 show support for the three predictions.

These findings have deep implications for ethical thinking: as Greene notes in the book (p. 251), they reverse the lesson philosophers thought we should derive from our answers to the footbridge dilemma.

The lesson is not that it is sometimes (as in footbridge) wrong to promote the greater good (saving five lives). It is that "our moral intuitions are generally sensible, but sometimes wrong." Greene suggests a hierarchy between the two modes of ethical thinking. This conclusion can be drawn because we are now understanding how our moral brains work. We have, in his words, found a guide for *moral psychology*.

5.3.1 *Shape of Things to Come*

Science might eventually take over the most interesting philosophical themes. But as research stands now, this is likely to occur in the distant future because progress in the experimental analysis of human nature is slow. One can look at the debate in the eighteenth century, as we are going to do next, as providing a rich set of hypotheses to test. We can compare, for instance, the well-known passage of the *Theory of Moral Sentiments* on how we perceive the blow to someone's leg[4] and the very similar experiment of the Fiery Cushman group reported in page 36 of Greene (2013) to see how this current new and exciting research is following the path indicated by philosophers. For our purposes, we will go back to the eighteenth century and rely on what they are still telling us. But let us start with the book at hand.

5.3.2 *An Illuminating Debate*

Very early in the book, the author reminds the reader of the exchange between Ron Paul and Wolf Blitzer reported in Greene (2013), page 7. In the exchange, Blitzer pressed Paul on the issue: what should "we" do when someone who freely decided not to buy health insurance finds himself in desperate need of medical care he cannot afford? After some skirmishes, Blitzer teased Paul: "But Congressman, are you saying that society should just let him die?" The author comments: "As Paul prepared his hesitant answer, a chorus of voices from the crowd shouted 'Yeah, let him die!' These are the Northern Herders." (The Northern Herders correspond in the book to modern-day conservatives; the Southern Herders the progressives.) Are the Northern Herders really so heartless? Here is a plausible alternative explanation of the "chorus of voices," (if I may lend them a voice):

"We are not taking Wolf Blitzer's bait. The discussion is not on what to do at the node in the extensive form game tree where the person who has not purchased the insurance falls sick, and is facing death. This is where Blitzer's finger is pointing, because he knows there is only one answer that anyone, even the most Northern of us Northern Herders, can give at that point. But we are not discussing here a moral dilemma, which is what takes place at that node. We are discussing policy, and the discussion concerns what to do when we design the outcomes of the game tree, with the clear understanding that if we write 'The insurance company will be forced to pay' at that node, then at the earlier node the agent will not buy the insurance. Instead if we write 'You will die,' the agent will buy the insurance. In neither case will the person die; the difference between the two cases is who pays for the health care, and whether someone is allowed to hitch a free ride. So by all means, let's write 'Let him die' on that node. This is not a matter of life and death, but of who pays for the health care."

The alternative explanation of "Yeah, let him die!" is, of course, that the Northern Herders do not see the game tree and are just heartless. We will never know what people in the audience really thought, but can we just jump to the conclusion that all those people miss the broader picture? An audience providing this articulate, well-reasoned answer

[4] "When we see a stroke aimed and just ready to fall upon the leg or arm of another person, we naturally shrink and draw back our own leg or our own arm" (Smith 1759, I.I.3).

would appear farsighted. On the other hand, the game is a simple two-player sequential move game, each player with two actions; this is as simple as it can get. If we think that no voter can see this, then we should doubt our own support for the democratic method of government.

A possible objection to our little speech is that the Northern Herders do not have a solution to the real problem, which is that long-term illness cannot be covered by private insurance contracts. This objection ignores economic analysis, which has long recognized that the Blitzer solution (let people not buy insurance, and then force the insurance company to offer a contract when long-term illness strikes) is not compatible with the existence of an insurance industry. Economic analysis has also recognized that long-term illness poses a problem for a market solution of the insurance problem (because the insurance has an incentive to cancel an existing contract in that event); but also that the solution cannot be long-term contracts (because the consumer cannot be forced into them). However, a solution of the problem is possible (Cochrane 1995): a sequence of time-consistent short-term contracts based on a series of state-contingent severance payments. This sequence constitutes a self-enforcing long-term contract and had been described at the time of the debate on the Clinton administration's health plan. So it may be the honest philosophical disagreement that the author mentions ("at the core this disagreement [of health care reform] is about the tension between individual rights and the (real or alleged) greater good" (Greene 2013)) is not necessary, if one understands basic tools of economic analysis.

The debate is a perfect illustration of the problems one runs into when the distinction between ethical reasoning and policy issues are not kept distinct. So the question is: which is the more realistic explanation, heartless or farsighted? This question illustrates the real dilemma facing economics today. This is the question we address next. In the next section, we present a preview of our main claims.

5.3.3 *A Preview of Our Claims*

We will claim that the correct way to view utilitarianism is as a constrained choice of the optimal constitution by a society, not as a guide for choice of actions by individuals; so it should be taken as providing a foundation for policy, not as an ethical doctrine. Of course the choice behind the veil of ignorance cannot be implemented by a real vote, but is an as-if construction (what would we choose if we could really choose behind the veil), so this poses the fundamental issue of how the policy choice is actually to be implemented.

We will claim that the answer one gives to the question of the optimal constitution should be based on a positive, "scientific" view of human nature, or, in the expression of Hutcheson (1725): "its Powers and Dispositions." Both elements are essential: *power* because we have to decide whether people are clever enough to see the future consequences of their actions; *dispositions*, because we have to decide which, on average, people's preferences are; in particular whether and how much they care about others. In very simple terms, are human beings good and stupid, as some of my behavioral economist friends seem to think, or selfish and clever?

We will claim that within the Scottish Enlightenment, two very distinct views of human nature were emerging, which had already appeared in the controversy between Bernard Mandeville and Jean Jacques Rousseau on the role they assigned to pity and self-love. Briefly, both agreed that pity is a fundamental motivation. The controversy centered on the role of self-love (or pride). In Mandeville, self-love is a passion that precedes society, and is the main engine of its development; in Rousseau, self-love follows

the establishment of society, creates inequality of outcomes, which in turn induces envy and resentment, which thus produces the corruption of society. We will claim that a proper balance between pity and pride is necessary for a complete understanding of human nature.

Finally, we will claim that far-sighted self-interest is just as important as the moral instinct in our social decision making.

6. *Utilitarianism Is Not a Religion*

Utilitarianism may be misunderstood as an ethical theory meant to provide guidance in individual choices, irrespective of what the other individuals in society are doing. Many of the applications Greene gives pertain to individual choice, and so he treats utilitarianism as a guide for individual conduct. This is a misunderstanding.

Let's see a typical example of this reasoning: "Thus says utilitarianism, you should spend the money helping desperately needy people rather than on luxuries for yourself" (Greene 2013, p. 207). The author does not seem to hesitate even when faced with the paradoxical conclusion of the *happiness pump*: "The utilitarian bleeding will continue until you've given away all of your disposable income." If utilitarianism is indeed interpreted as a moral command on individual behavior, then the conclusion follows from diminishing marginal utility. More formally, if the command is: "Maximize by means of unilateral transfers the sum of the utilities of all the people in the world, including yours, under the budget constraint that you cannot transfer more than what you have and a non-negativity constraint on transfers (you cannot rob others)," then indeed you should reduce your current income to that of the poorest people in the world. The author has sufficient sense of realism to realize that this is not a feasible proposal; but he feels guilty about this. He is not willing to forgive the human species for its limitations, and for not living up to the ideal, either. The best option we have is to humbly acknowledge our hypocrisy: "Speaking for myself, I spend money on my children that would be better spent on distant starving children, and I have no intention of stopping. After all, I am only human! But I would rather be a human who knows he is a hypocrite, and who tries to be less so, than one who mistakes his species-typical moral limitations for ideal values" (Greene 2013, p. 268).

But is this really what utilitarianism prescribes? This misunderstanding is likely to occur because it is pervasive in the presentation of the founding fathers of Utilitarianism, in particular J. S. Mill's. Consider, for instance, the formal statement Mill provides early on in his treatise on the subject: "The creed which accepts as foundation of morals, Utility, or the Greatest Happiness Principle, holds that actions are right in proportion as they tend to promote happiness, wrong as they tend to produce the reverse of happiness. By happiness is intended pleasure, and the absence of pain; by unhappiness, pain, and the privation of pleasure" (Mill 1863, ch. II, "What Utilitarianism Is").

That there is confusion was made clear by an attempt to introduce clarity with the distinction between act utilitarianism and rule utilitarianism. The terminology was introduced in the late 1950s by Richard Brandt (Brandt 1959). Interestingly, John Rawls (1955) was an early forerunner of rule utilitarianism, approximately at the time in which John Harsanyi was developing his application to social welfare of expected utility. *Act utilitarianism* requires the greatest happiness principle to be applied separately for each action, just as in the statement by Mill we reported. In the choice between two actions, act utilitarianism requires us to choose the one that maximizes the sum of the utilities of all men in the community.

All men must be included by the impartiality clause.[5] *Rule utilitarianism* requires first that rules have to be established that guarantee the greatest happiness principle over all the foreseeable future applications to specific action choice; then each action has to be evaluated not on the basis of the happiness it promotes in the specific instance, but on the consistency with those rules. It is known that the definition given by Mill that we reported, that appears as the textbook definition of act utilitarianism, is then followed, a few pages later, by what appears as the textbook definition of rule utilitarianism: "In the case of abstinences indeed—of things which people forbear to do from moral considerations, though the consequences in the particular case might be beneficial—it would be unworthy of an intelligent agent not to be consciously aware that the action is of a class which, if practiced generally, would be generally injurious, and that this is the ground of the obligation to abstain from it" (Mill 1863). So the confusion was real.

The distinction between act and rule utilitarianism is now outdated. The solution of what utilitarianism should be interpreted is provided in Harsanyi (1953, 1955) and Rawls (1971, 2001) by specifying two basic propositions. The first proposition is that we choose among rules, not individual actions, on the basis of our preferences. The correct inference from utilitarian premises is not an ethical conclusion, but a constitutional one, not what is right to do, but what social procedures we agree on. "Value judgments concerning social welfare are a special class of judgments of preference, inasmuch as they are nonegoistic impersonal judgments of preference" (Harsanyi 1953, p. 434). The second proposition is that in deciding rules, any personal interest has to be excluded. "A value judgment on the distribution of income would show the required impersonality to the highest degree if the person who made this judgment had to choose a particular income distribution in complete ignorance of what his own relative position (and the position of those near to his heart) would be within the system chosen" (Harsanyi 1953, p. 435).

Even this version of utilitarianism (or rule utilitarianism) does not provide the precise prescription of policy that perhaps Greene is hoping for. The Pareto weights that can be assigned to each individual are not unique, of course. There is even a discussion currently on who is an individual in this maximization. For example, in a two-periods society where father and children can be of two types (say high and low skill), is the child a single individual, or rather two, one for each type realization? The usual answer is one; the more natural answer, we argue, is two (Phelan and Rustichini forthcoming)—and this second view reduces the need to provide insurance for the hypothetical person "child." Greene is in good company in his hope that utilitarianism can provide a precise, unique answer to social-policy problems: Mill had a similar opinion. In chapter 5 of *Utilitarianism* (Mill 1863), he argues that the greatest happiness principle can provide criteria that are consistent with, and further restrict, those provided by justice: "Social utility alone can decide the preference" he says, for instance, at the conclusion of the analysis of the problem of income distribution. We disagree. Modern taxation theory in a standard Mirrlees environment cannot even decide in favor of progressive taxation unless one introduces a "taste for equality" (Fahri and Werning 2010). Similar considerations hold in the case of abortion. Greene emphasizes that different gains and losses should be considered, and once this is done, an argument overall in favor of only minor restrictions on abortion seems to appear. A pro-life utilitarian might follow the same method, put more emphasis on the opportunities provided by

[5] This is "Bentham's dictum, *everybody to count for one, nobody for more than one*," Mill (1863), chapter 5.

birth control, and thus reach a very different conclusion.

7. *Science of Man*

In our discussion of the doctrine of double effect, we noted that in looking for the moral foundation of laws, and specifically on whether killing in self-defense is admissible, Aquinas looked for intentions as the benchmark; and we wondered why intentions should have this role in his system. The reason is that the purpose of the laws is to remove the cause of evil, and the only way to do this is modify the behavior. Consider the following statements:

> Moral or ethical virtue is the product of habit and has indeed derived his name [. . .] from that word (Aristotle 1926, Book II, 1). Virtues are dispositions (Aristotle 1926, Book II, v, 3-vi). The dispositions are the formed states of character in virtue of which we are well or ill-disposed in respect of the emotions (Aristotle 1926, Book II, iv, 4-v).

> [. . .] human virtue which is an operative habit, is a good habit, productive of good works (Aquinas 1265–73, I-II, q.55, a.3).

> men never work any good unless through necessity [. . .] Therefore it is said that hunger and poverty make men industrious, and the laws make them good (Machiavelli 1531, chapter 3).

> To breed an animal who makes promises (versprechen darf)—Is this not the paradoxical task nature has set itself with respect to humans? (Nietzsche 1887).

The trait common to these quotes is clear and indicates the reason for Aquinas's emphasis on intentions. Ethics and policy induce behavioral modification; intentions are important because the purpose of the law is to change behavior by changing intentions: *in interiore homine habitat veritas*. But how far can behavior modification go? On this question, the research presented in Greene (2013) will have a fundamental role in future research. We have seen that regions in the brain code the utilitarian computation. What is the implication of the finding that the neural basis of some norms or computations can be traced to brain structures? Utilitarianism has to set up rules that maximize utility given the constraints provided by power and dispositions, that is by the structure of preferences and cognitive abilities of the people that have then to follow the rules. So the biological structures provide the constraints for rules setting. Understanding these constraints is the key to avoiding utopian rules setting.

For example, suppose a researcher did try to run an experiment where the subject in the scanner is asked to choose in the *Most* dilemma. So he is facing the bystander at the switch dilemma, but the single person is his son. Would the pattern of brain activity be different from the standard bystander-at-the-switch case? Most certainly it would. Should we take this into account? Of course we should. We are not proposing with this to derive an "ought" from an "is": we are instead recognizing that the design of an optimal mechanism, or constitution, has to be based on a clear idea of what the feasible set is, and what the constraints are. Human nature is the most important of these constraints. In contrast, the author of *Moral Tribes* has a higher aspiration for us, and even if evolution has worked to produce a set of psychological adaptations called morality, a more perfect species is desirable, free of the "species-typical moral limitations" (Greene 2013, p. 268). The book may be seen as a blueprint for this enterprise.

The idea we indicate is already present in these early authors, in Mandeville "Be fam'd in War, yet live in Ease, Without great Vices, is a vain EUTOPIA seated in the Brain" (Mandeville 1705–29, *The Grumbling Hive*), as in Hutcheson (1725): "There is no part of

Philosophy of more importance, than a just knowledge of Human Nature, and its various Powers and Dispositions" (*The Preface*). David Hume elevated these observations into a methodology.[6]

In laying the foundation of the science of man, two lines of thought developed. It is time to review them.

8. *The Great Divide*

> I have received, sir, your new book against the human species [. . .] Never so much cleverness has been put to use trying to turn us to brutes: when one reads your book he is taken by the desire of walking on all fours (Voltaire, Letter to Rousseau, August 30, 1755).

> The second volume of the Fable of the Bees has given occasion to the system of Mr. Rousseau, in whom however the principles of the English author are softened, improved and embellished, and stript of all that tendency to corruption and licentiousness which has disgraced them in the original author (Adam Smith, Letter to the Authors of the *Edinburg Review*, section 10, published 1756).

In the summer of 1755, two letters, written at the same time by two eminent thinkers, commented on the recent book by the man of Geneva. Reading the two letters side by side, one is left wondering who was the economist and who was just the philosopher.

In the first letter, the grumpy old (sixty-one years) French philosopher goes straight to the heart of Rousseau's philosophy and its profoundly reactionary thesis that society leads to corruption of the innocence in the state of nature. "[T]he example of savages, most of whom have been found in this state, seems to prove that men were meant to remain in it, it is the real youth of the world, and that all subsequent advances have been apparently so many steps toward the perfection of the individual, but in reality towards the decrepitude of the species."[7] Voltaire was no friend of power and authority. He had just been arrested by his former employer, Frederick the Great, who had then proceeded to burn all the copies of his attack on Pierre Louis Maupertuis. He had then taken refuge in Paris, where he was banned, this time by Louis XV, and had finally taken refuge near Geneva. His notes at the margin of his copy of the *Discours* show that he had read it carefully, and even more critically than the abrasive tone of the letter reveals.

The second is the "Letter to the Authors of the *Edinburgh Review*," by Adam Smith. The letter is a report that the eager young (thirty-two years) scholar Smith sent to the *Review* to update the audience on the most recent philosophical developments in continental philosophy. In the substantive part of the letter, Smith attempts the implausible maneuver of reconciling Mandeville of the *Fable* with Rousseau of the *Discours*, presenting the latter as the only possible rendition of the first that one could possibly present in polite company. The *Treatise on Moral Sentiments*, published just four years later, is the opus that Smith intended as a reconciliation of Mandeville and Rousseau.

But what was the root of the dissent between Mandeville and Rousseau? The method they use is the same (as Smith noted in his letter). Both appeal to the conceptual device of the state of nature. This is an exercise in abstraction: man in the state of nature is what is left once we remove all that the

[6] "Here then is the only expedient, from which we can hope for success in our philosophical researches [. . .] to march up directly to the capital or center of these sciences, to human nature itself; which being once masters of, we may everywhere else hope for an easy victory. [. . .] There is no question of importance, whose decision is not compriz'd in the science of man" (Hume 1739–40, introduction).

[7] L'exemple des sauvages qu'on a presque tous trouvés à ce point semble confirmer que le genre humain ètait fait pour y rester toujours, que cet état est la veritable jeunesse du monde et que tous les progrès ultérieurs ont été en apparence autant de pas vers la perfection de l'individu, et en effet vers la décrépitude de l'espèce (Rousseau 1755).

social environment adds. In the words of Rousseau: "If we strip this being, thus constituted, of all the supernatural gifts he may have received, and all the artificial faculties he can have acquired only by a long process; if we consider him, in a word, just as he must have come from the hands of nature. . ." In this sense, state of nature is conceptually equivalent to what is genetically determined, to nature as opposed to nurture.

Both authors also agree on the fundamental components of human nature. For both authors, the components are two, and the definitions they provide of these components are very similar. The pity of Mandeville is *pitié* of Rousseau, and self-love (or self-liking) of Mandeville is *l'amour de soi-même* of Rousseau. They are both fundamental.[8] The controversy between Mandeville and Rousseau appears to be (but we will claim that this is not the crucial difference) essentially about the nature of virtue, and whether pity qualifies as a virtue. For Mandeville, virtue requires completely unselfish motivations. "The generality of moralists and Philosophers have hitherto agreed that there could be no Virtue without Self-denial" (Mandeville 1705–29, *An Essay on Charity and Charity Schools*). As a corollary, pity is a passion, not a virtue. In contrast, for Mandeville charity is a virtue: "This virtue [charity] is often counterfeited by a passion of ours, call'd Pity or Compassions". For Rousseau, instead, pity is the source of all social virtues.[9]

Mandeville thought this to be a cornerstone of his system: pity is really not benevolent, but ultimately driven by self-love. Everyone, especially his critics, passionately agreed with him, because they thought this would disqualify his entire system into shame. Mandeville was aware of the risk implicit in this position, and he devoted the second volume of the *Fable* to a systematic response to Joseph Butler's objections to the first volume. Butler's objections to Mandeville were targeted on psychological egoism; namely the idea that all actions, even those that appear motivated by love of others, are in fact selfish, because all actions are motivated by the pleasure that one derives from them. This debate is illuminating for economists because economists are Mandevillian by construction. Psychological egoism; namely the idea that all actions, even those that appear motivated by love of others, are in fact selfish, because motivated by the pleasure that one derives from them. This follows from the way we construct utility. Economists observe an individual selecting one option out of each feasible set, and assume that enough technical conditions hold so that the selection can be represented as maximization of a utility function. Thus, human choice in economics is, by construction, the result of selfish maximization. Butler's now classical argument ("Butler's stone") is an argument against this method. He notes that enjoyment follows from a specific property of the object: in his example, food, as opposed to a stone.[10] Thus, it is not pleasure that we are looking for, but those specific properties; similarly, when we pursue the benefit of others, it is this benefit (just like the food), not our own pleasure in helping others, that we desire. Butler

[8] "La pitié est un sentiment naturel qui, modérant dans chaque individu l'activité de l'amour de soi-même, concourt à la conservation mutuelle de toute l'espèce" (Rousseau 1755).

[9] Rousseau noted clearly the difference: "Mandeville well knew that, in spite of all their morality, men would have never been better than monsters, had not nature bestowed on them a sense of compassion, to aid their reason: but he did not see that from this quality alone flow all those social virtues, of which he denied man the possession" (Rousseau 1755).

[10] "That all particular appetites and passions are towards external things themselves, distinct from the pleasure arising from them, is manifested from hence; that there could not be this pleasure, were it not for that prior suitableness between the object and the passion: there could be no enjoyment or delight from one thing more than another, from eating food more than from swallowing a stone, if there were not an affection or appetite to one thing more than another" (Butler 1726, Sermon XI, p. 365).

is forcing Mandeville and economists out of a comfortable defensive position ("We are psychological egoists because that is the definition of utility maximization"). The real question is: what specific properties do you think it is natural for humans to pursue, those that possess Butler's "prior suitableness"? To this question we have to give a substantive answer.

Mandeville took Butler's challenge. In his reformulation, self-love is the passion motivating individuals pursuing their utility; self-liking is pride, the specific property that Butler was asking him to exhibit. He then observed that pride cannot be affirmed only by one's own approval, as obviously this has very little power of proof of the quality of an individual. Instead, it requires the approval of others, and this is what motivates us to look for it. So Mandeville separates the consequence (others' approval and consequent self-liking) from utility. Thus, acts that favor others are seen as ways to procure others' approval, and so provide us with justified self-liking; they have the property of the food, rather than the stone, in Butler's fundamental dichotomy. In this way, Mandeville goes beyond psychological egoism as a formal property ("we are selfish because by definition we maximize our utility, even when we sacrifice all we have to others") to a substantial property.

Rousseau addresses the same issue and concludes that self-love is absent in the state of nature, and only induced in man by society. The "inequality of credit and authority became unavoidable among private persons, as soon as their union in a single society made them compare themselves one with another and take into account the differences which they found out from the continual intercourse every man had to have with his neighbors."[11] In Mandeville, self-love is the condition for the existence of a functioning society. This is, after all, the moral of the poem: replacing selfishness with honesty produces stagnation.

Coherently with the *Letter to the Edinburgh Review*, Smith criticized Mandeville and devoted a special attack to his system (Smith 1759, Part VII, section II, chapter IV). The verdict was harsh: "the notions of this author are in almost every respect erroneous." Smith was a respectable philosopher, and no respectable philosopher would have liked to be seen in the company of Mandeville.[12] Smith's main argument was against the identification of self-love with vanity, and the general idea that self-denial is not a necessary condition for virtue. Interestingly, however, Smith noted that it was an easy task for Mandeville (in his polemic with "some popular ascetic doctrines," perhaps Shaftesbury's) that "the extirpation and annihilation of all our passions [. . .] would be pernicious to society, by putting an end to all industry and commerce." So on the crucial issue, Adam Smith is a follower of Mandeville, and this would become one of the dominant themes of *The Wealth of Nations*.

A superficial reading of *The Theory of Moral Sentiments* may suggest that, since sympathy appears as the fundamental link among men, a natural generosity will spring from it, and balance the selfishness of *Homo economicus*. The opposite is true. This should be clear by the simple observation that in *The Wealth of Nations*, when Smith examines the conditions for an efficient cooperation,

[11] l'inégalité de crédit et d'autorité devient inévitable entre les particuliers sitôt que réunis en une même société ils sont forcés de se comparer entre eux et de tenir compte des différences qu'ils trouvent dans l'usage continuel qu'ils ont à faire les uns des autres.

[12] Adam Smith was a very cautious man: see for example Lomonaco (2002), p. 672; and the *Fable* was a book of ill repute, as it was " convicted as a nuisance by the grand jury of Middlesex in 1723, which stands out in the history of the moral sciences for its scandalous reputation. Only one man is recorded as having spoken a good word for it, namely Dr. Johnson" (Keynes 1936, chapter 23, section VII).

sympathy is never mentioned and never called as a condition for the smooth working of a market society. This absence of sympathy from the treatise on economics may be explained by the weakness of sympathy when compared to self-interest. There are well-known passages in Smith (1759) that are usually cited to justify this interpretation, in particular the famous paragraph on the European who would not sacrifice the little finger for one hundred million Chinese lives (Book 1, part III, chapter III). Greene (2013, p. 201) correctly points out that we recognize this thought to be monstrous, and we reject it. But a careful reading of Smith shows that more than a utilitarian calculus, based on the acceptance of an abstract impartiality rule, it is the desire of approval of others, which is the foundation of the sense of duty that moves us in this case. There is a deeper reason for why Smith did not see sympathy as the antidote to self-interest, and as the asymmetry of sympathy, not its weakness. He (Smith 1759, Book 1, part 1, section III, chapter II, "On the origin of ambition and the distinction of rank") notes that naturally mankind sympathizes more with joy than with sorrow. Understanding this, and anticipating this response of others, we then devote our energies to avoid the humiliating situation where our distress is open to the eyes of the public, and we are aware that "no mortal conceives for us the half of what we suffer." Thus "it is chiefly from this regard to the sentiments of mankind that we pursue riches and avoid poverty."

As Greene clearly demonstrates in his book, we now have the tools to produce a science of man based on a broad evidence obtained by borrowing tools from psychology, economics, neuroscience, and genetics. But this research will have to produce a complete model of human nature. Emotions such as envy, regret, blame, and the pleasure for competitive achievement are notably absent. Similarly, the role of hypocrisy has been barely scratched. There are some exceptions (for example, Dana, Weber, and Kuang 2006, Trivers 2011, Bault et al. 2011, Gurdal, Miller, and Rustichini 2013), but they are a small fraction compared to the large body of research on altruism. If we lack this complete model, our policy choices attempting to implement the utilitarian maximization in the choice behind the veil will be based on wrong premises.

9. *Far-Sighted Agents*

Let us go back to the Blitzer–Paul debate, and the crucial question of the farsightedness of that audience, or more generally of mankind. The method of choice behind the veil of ignorance (and the two propositions that constitute it) is a premise implicitly adopted in the trolley literature. Obviously, the non-egoistic condition is satisfied in each single dilemma, as the decision maker has no personal interest in the outcome. A single potential exception to this rule is in a twist to "transplant" on page 1399 of Thomson (1985). In this example, patients need the body parts because the surgeon was careless in an earlier intervention; thus the surgeon has a direct interest in the choice: but it is far from clear whether the surgeon is also the decision maker. Second, less obviously, subjects are made to choose on a very specific episode, with carefully delimited consequences, for a good reason. As an experimental device to test hypotheses about moral psychology, the trolley methodology consists in gathering evidence on what principles individuals use when they make moral judgments. The idea is that if we ask parable-like questions, the subject will unconsciously use general principles. In this way, we will cleverly read their mind, and even scan their brains, to flesh these principles out. The subject is not invited to, and is assumed not to, consider general principles. How can we assume this? How do we know they do not?

More likely, once again the experimenter has lost control of the experiment.

The class of trolley problems are a large exercise in the application of the categorical imperative in the *first*, more familiar formulation ("Act only on that maxim through which you can at the same time will that it should become a universal law"). This is precisely the incentive compatibility condition in the choice behind the veil of ignorance, although the foundation (as the illustrations Immanuel Kant gives) is not based on preferences, but on the clause that the maxim should not contradict the purpose of the institution. For example, I cannot borrow money promising to pay it back with the intention of not keeping the promise; that would make "the promise and the intended purpose of it impossible."

There is only one way to find out what subjects really think, and it is the experimental method; but the hypothesis should be tested with a different experimental design. Suppose that the question in "transplant" was explicitly framed as a juridical norm. That is, in the transplant dilemma, rather than asking as Thomson suggests—"Would it be morally permissible for you to operate anyway?"—we should ask, "Would you be in favor of a legal provision that allows truly great surgeons to choose a patient in the hospital and use his body to farm organs for other patients, provided the number of patients gained versus lost is greater than or equal to five?" Just as in the case of the debate between Blitzer and Paul, the questions we should address here are: Are subjects who think about the moral dilemmas aware of the implications of the policy to which they are subscribing? Can they see the wider extensive-form game, or are they just looking at the node that the experimenter is pointing at? Are they taking the experimenter's bait? The implications of a choice may be hidden to the chooser; maybe they are established not by explicit acceptance, but by setting a precedent.

The pattern of brain activation would help to test which of these hypotheses is true. For example, how do we know that activation in dorsolateral prefrontal cortex is just evidence of the adding and subtracting that are involved in the utilitarian calculation, and not as well the consideration of the wider implications of an action? That is, why should we think that activity is only Bentham's (Bentham 1948) and not also Kant's?

We have strong reasons to believe that far sightedness is as important as moral emotions in human interactions. Recent research on the role of intelligence in influencing strategic behavior shows that an important, and possibly the main, explanatory factor of cooperative behavior is intelligence (Rustichini 2015); its predictive power is stronger than any of the other traits, like conscientiousness and agreeableness.

We have now the conceptual and technical conceptual means to produce a science of man going beyond the imagination of the philosophers who set up the terms of the debate on ethics. On the basis of the premise that ethics and policy to be effective have to be constrained by the characteristics of human nature we have to produce this science. This includes more than preferences, such as discount and risk aversion, as economics has done so far, to include "powers and dispositions." If this is the future program, the portrait that is given in *Moral Tribes* is illuminating and fascinating, but incomplete.

References

Anscombe, G. E. M. 1967. "Who Is Wronged? Philippa Foot on Double Effect: One Point." *Oxford Review* 5: 16–17.

Aquinas, Thomas. 1265–1273. *Summa Theologae*.

Aristotle. 1926. *Nicomachean Ethics*. Cambridge: Harvard University Press.

Bault, Nadège, Mateus Joffily, Aldo Rustichini, and Giorgio Coricelli. 2011. "Medial Prefrontal Cortex and Striatum Mediate the Influence of Social Comparison on the Decision Process." *Proceedings of the National Academy of Sciences* 108 (38): 16044–49.

Bentham, Jeremy. 1948. An Introduction to the Principles of Morals and Legislation. New York: Hafner

Publishing Co.
Bonnefon, Jean-François, Azim Shariff, and Iyad Rahwan. 2015. "Autonomous Vehicles Need Experimental Ethics: Are We Ready for Utilitarian Cars?" Research Forum on Semi-autonomous Systems: Trust, Control, and Letting Go, Melbourne, Australia, November 10–11.
Brandt, Richard B. 1959. *Ethical Theory: The Problems of Normative and Critical Ethics*. Englewood Cliffs: Prentice-Hall.
Butler, Joseph. 1726. *Fifteen Sermons Preached at the Rolls Chapel*. London: J. and J. Knapton.
Cochrane, John H. 1995. "Time-Consistent Health Insurance." *Journal of Political Economy* 103 (3): 445–73.
Dana, Jason, Roberto A. Weber, and Jason Xi Kuang. 2007. "Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness." *Economic Theory* 33 (1): 67–80.
Farhi, Emmanuel, and Iván Werning. 2010. "Progressive Estate Taxation." *Quarterly Journal of Economics* 125 (2): 635–73.
Foot, Philippa. 1967. "The Problem of Abortion and the Doctrine of Double Effect." *Oxford Review* 5: 5–15.
Greene, Joshua D. 2013. *Moral Tribes: Emotion, Reason, and the Gap between Us and Them*. London: Penguin Random House.
Greene, Joshua D., and Jonathan Haidt. 2002. "How (and Where) Does Moral Judgment Work?" *Trends in Cognitive Sciences* 6 (12): 517–23.
Greene, Joshua D., Leigh E. Nystrom, Andrew D. Engell, John M. Darley, and Jonathan D. Cohen. 2004. "The Neural Bases of Cognitive Conflict and Control in Moral Judgment." *Neuron* 44 (2): 389–400.
Greene, Joshua D., R. Brian Sommerville, Leigh E. Nystrom, John M. Darley, and Jonathan D. Cohen. 2001. "An fMRI Investigation of Emotional Engagement in Moral Judgment." *Science* 293 (5537): 2105–08.
Gurdal, Mehmet Y., Joshua B. Miller, and Aldo Rustichini. 2013. "Why Blame?" *Journal of Political Economy* 121 (6): 1205–47.
Harsanyi, John C. 1953. "Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking." *Journal of Political Economy* 61 (5): 434–35.
Harsanyi, John C. 1955. "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility." *Journal of Political Economy* 63 (4): 309–21.
Hume, David. 1739–40. *A Treatise of Human Nature*. London: John Noon.
Hutcheson, Francis. 1725. *An Inquiry into the Original of Our Ideas of Beauty and Virtue*. Indianapolis: Liberty Fund, 2008.
Kant, Immanuel. 1785. *Groundwork for the Metaphysic of Morals*, edited by Mary Gregor and Jens Timmermann, 1998.
Keynes, John M. 1936. *The General Theory of Employment, Interest and Money*. London: Palgrave Macmillan.
Lomonaco, Jeffrey. 2002. "Adam Smith's 'Letter to the Authors of the *Edinburgh Review*'." *Journal of the History of Ideas* 63 (4): 659–76.
Machiavelli, Niccolò. 1531. *Discorsi sopra la prima deca di Tito Livio*. Firenze: Sansoni, 1971.
Mandeville, Bernard. 1705–29. *The Fable of the Bees, or Private Vices, Publick Benefits*. New York: Penguins Classics, 1989.
Mill, John Stuart. 1863. *Utilitarianism*. London: Parker, Son and Bourne.
Nietzsche, Friedrich. 1887. *On the Genealogy of Morals*, edited by Walter Kaufmann. New York: Random House, 1989.
Phelan, Christopher, and Aldo Rustichini. Forthcoming. "Pareto Efficiency and Identity." *Theoretical Economics*.
Rawls, John. 1955. "Two Concepts of Rules." *Philosophical Review* 64 (1): 3–32.
Rawls, John. 1971. *A Theory of Justice*. Cambridge: Harvard University, Belknap Press.
Rawls, John. 2001. *Justice as Fairness: A Restatement*, edited by Erin Kelly. Cambridge: Harvard University, Belknap Press.
Rousseau, Jean-Jacques. 1755. *Discourse on the Origins and Foundations of the Inequality among Men*, edited by Maurice Cranston. London: Penguin Classics, 1984.
Rustichini, Aldo. 2015. "The Role of Intelligence in Economic Decision Making." *Current Opinion in Behavioral Sciences* 5: 32–36.
Smith, Adam. 1759. *The Theory of Moral Sentiments*. New York: Penguin, 2010.
Thomson, Judith Jarvis. 1985. "The Trolley Problem." *Yale Law Journal* 94 (6): 1395–415.
Trivers, Robert. 2011. *The Folly of Fools: The Logic of Deceit and Self-Deception in Human Life*. New York: Basic Books.
Voltaire. 1755. "Letter to J. J. Rousseau." In *Voltaire in His Letters, Being a Selection from His Correspondence*, edited by S. G. Tallentyre, 146–53. New York and London: G. P. Putnam's Sons, 1919.