Rawls's "original position" is not sufficient to specify the rules of cooperation.

Bahram Houchmandzadeh

Rawls's "original position" is not sufficient to specify the rules of cooperations.

Bahram Houchmandzadeh
CNRS, LIPHY, F-38000 Grenoble, France
Univ. Grenoble Alpes, LIPHY,
F-38000 Grenoble, France

In his landmark work, "justice as fairness", John Rawls conjectured that the concept of "original position" and "veil of ignorance" will specify the terms of cooperation the individuals forming a society would agree to. We show here that this concept does not carry enough constraints and therefore is not sufficient to reach such an agreement.

I. INTRODUCTION.

John Rawls, in his landmark work, "Theory of justice" [1, 2], introduced the concepts of "Veil of ignorance" and the "original position" into the contractarianism tradition of political philosophy[3]. He argued that the *veil* strips away all biases from the contractors - due to their sexes, origin, religion,... - during their deliberation (figure 1). Therefore, he conjectured that "the principles of justice the parties would agree to [...] would specify the terms of [their] cooperation"[2, p. 17]. Rawls then developed this idea and concluded that the principle of justice the parties would agree to would be the so called Maximin, a set of rules that maximizes the outcome of the worst-off person of the society.

From its onset, the Maximin deduction came upon criticisms from the *utilitarians* [4, 5] which favor maximizing the total outcome ; various other principles between Maximin and utilitarian can be argued about[6], depending on the weight given to the risk adversity of the rational human.

The purpose of this article is to argue that the debates between these various schools is beside the point : even if the parties, using the veil of ignorance, agree to the same principle of justice (Maximin, utilitarian,...) they cannot agree on a set of rules to achieve the desired principle. In short, the "veil of ignorance" does not remove enough of the original biases to allow for such an agreement.

In order to demonstrate the above statement, I will use a simple mathematical formulation of Rawls theory. Usually, most areas of philosophy are too complex to be modelizable by a mathematical approach. The theory of Rawls however shares many features - such as principles of symmetry, use of initial conditions, optimization procedure, ... - with physical theories. The theory has been restricted (to the most basic social institutions) and abstracted (Rawls insists many times that his view is a theoretical tool and does not reflect any real situation) and all of the simplifying hypothesis have been precisely defined. All these features allows one to put the theory to a mathematical test and this is precisely the purpose of this article.

It is of course not the first time that such an attempt has been made. The qualitative mathematical formalism was used by Rawls himself when discussing his second principle of justice and introducing the difference principle (maximin). The features I have enumerated above have attracted many scientists to mathematical modeling (see [7] for a review) but, to my knowledge, these works have mostly been dedicated to decide which principles (maximin, utilitarian, ...) can be deduced from the "original position" hypothesis.

In the next section, I introduce a simple mathematical modeling of the Rawls theory. This model is restricted to a minimalistic version of Rawls theory, where the outcome for each individual (his quality of life) in the society can be measured by a single number (such as his wealth). I'll show that even in this simplest model, no agreement can be reached and therefore, any more realistic model will only increase the amount of indeterminacy.

The details of mathematical demonstrations are given in the appendices.

II. MINIMALISTIC RAWLS .

A. Veil of ignorance.

The minimalist version of Rawls theory I use is the following (figure 1) : N individuals decide to join and form a society, where the relations between individuals obey a given rule $\mathcal{R}$ (Slavery, Feudalism, Capitalism, Communism, ...) . In particular, the choice of $\mathcal{R}$ will decide how wealth and more generally "quality of life" will be distributed in the society. I suppose that this "quality of life" can be measured by a single number. We can assume that when individuals confer on such matters, their selection of the rule $\mathcal{R}$ is biased by their current position in the present society : A wealthy person will argue for the superiority of capitalism and the sanctity of private property, a feudal lord will be surprised to hear about egalitarian society while a leather worker of 1830 would adhere to the thesis of Fourier, St-Simon, or Proudhon.

Is there a possibility to choose a "*just*" rule $\mathcal{R}$? Can such a thing even exist ? Rawls found an ingenious device to solve this dilemma ("a theory of justice", 1971[1]) in the concept of "Veil of Ignorance". The society the individuals will join is made of vacant positions (for feudalism : few lords, a little more miserable merchants, a large number of totally miserable serfs), and each individual will pick *at random* its new position is the chosen society. In particular, an individual will not know his

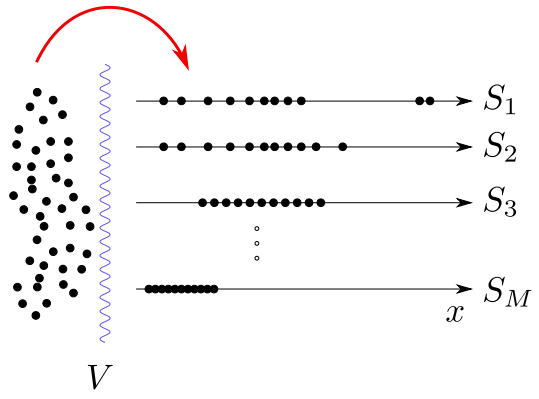

Figure 1. The abstract concept of "social contract" and the formation of society from free individuals. Individuals consent to a rule $\mathcal{R}_i$ which will result in a type of society S_i they will join, when a certain kind of resource distribution has been agreed upon. The idea of Rawls is that free individuals are behind a "Veil of ignorance", not knowing who they are going to be (gender, skin color, education, ...) in the society the will join : they will pick at *random* a place in the society they have chosen.

future age, gender, color, education level. If the chosen society is feudalism and he has picked the position of a female serf, he *will* become an authentic female serf with all the attributes and cognitive capital of such a person in such a society. His brain will be totally rewired and he will loose all memory of his initial position.

Individuals behind the veil will consider all possible societies S_1, ...,S_M and choose the *best* one among them.

Does the notion of Veil of ignorance allows one to decide which is best ? A short summary of Rawls's answer is "yes" and the solution is the "Maximin" solution : individuals would not want to be in a very bad position in the new society ; in order to limit such a risk, they will choose an organization S_α which maximizes the wealth of the least favored person.

The Maximin principle solution is the solution if we consider humans to be totally risk averse. The utilitarians would argue that a rational human should maximize its "expectation" and would therefore choose a society that maximizes the total wealth. Mixing these two strategies (avoiding risk and maximizing expectation) by giving various weight to each factor, depending on our conception of human behavior, will lead to different criterion for choosing the best society.

B. Indeterminacy of the veil.

The problem with the veil of ignorance is that individuals behind it cannot truly *observe* the possible societies and gauge them. The only thing they can do is to *predict* the kind of society they will create *if* they agree on a set of rules. These predictions are just that : predictions. They depend on the theoretical models in which the individuals believe. Therefore, if the individuals behind the veil have different theoretical models, even if they believe in the same principle of justice, they cannot agree on a set of rules to achieve this principle.

For example, suppose that all individuals behind the veil adhere to the Maximin as the principle of justice. Some individuals, let us call them neutralist, believe that the total output of the society (or its total wealth) depends mainly on the number of its members and is not too much affected by their inner relations. With this theoretical model in mind, they would conclude that the society with the best Maximin rating would be a totally egalitarian one. Accordingly, they will choose a set of rules $\mathcal{R}_1$ that is completely redistributive (the mathematical proof of these statements are given in the next section) such as a Marxist one or one with a very progressive taxation scheme.

On the other hand, some individuals believe that inequality among the society members will enhance competition and the total output of wealth. Let us call this belief the "trickle-down" theory. With this theoretical model in mind, they would conclude that the society with the best Maximin rating would be a totally inequalitarian order. Indeed, if inequality creates enough excess wealth (compared to an egalitarian society), the worst-off person in their ideal society would have superior wealth compared to the worst-off person in an egalitarian one. Trickle down people therefore would agree to a set of rules $\mathcal{R}_2$ that enhances capital accumulation feedback and let 1% of the society (or less) possess most of the total wealth.

Let me stress this point : in the above textbook case, both the neutralist and the "trickle-down" believe in the Maximin principle of justice. However, the rules they would accept to achieve this principle are diametrically opposite.

The veil of ignorance removes all information about the status of an individual in the future, after the veil, but it does not remove his present state of mind and the theoretical models he adheres to. Even if all individuals adhere to the same principle of justice, variability in their inner theoretical models would prevent them to accept the same set of rules.

In the following, using a very simple mathematical modeling, I will demonstrate the above statements. I stress that I consider only very simple models as "proof of principle". Adding more realism and complexity to the mode only enhances the indeterminacy of the Veil.

III. MATHEMATICAL FORMULATION.

In order to choose a "best" society, we need a comparison tool (a weighting tool) to numerically order the available ones. Let us call x the abstract measure of quality of life. Von Neumann and Morgenstein [8] introduced such quantification in 1944 when modeling the game theory ; in a more restricted sense, x would represent the

wealth of individuals. A society S_i is characterized by the value of its vacant positions $(x_1, x_2, ..., x_N)$ (figure 1) which we will denote by $S_i(\mathbf{x})$. The *function* $S_i(\mathbf{x})$ captures entirely the distribution of x for society i. For example, $S_1(\mathbf{x}) = \{1, 2, 4\}$ designates a society of 3 individuals where the worst-off one has wealth equal to 1 unit ($x_1 = 1$) while the best-off one has wealth equal to 4 unit, ; $S_2(\mathbf{x}) = \{0.5, 1, 6.5\}$ designates a less egalitarian society where the wealth of the worst-off person is only 0.5 unit. The convention in this article is to use ordered set and x_1 designates always the wealth of the worst-off society.

Once we have characterized the various societies by their functions, we can choose a criterion W to ascertain their merit. In other words, $W[S_i(\mathbf{x})]$ associates a single number (a merit) to society i with wealth distribution $S_i(\mathbf{x})$. For example, $W_R[S(\mathbf{x})] = \min\{S(\mathbf{x})\}$ uses the wealth of the worst-off person to gauge each society. In the above examples, $W_R[S_1(\mathbf{x})] = 1$, $W_R[S_2(\mathbf{x})] = 0.5$ and therefore, in the light of *this* merit functional, society S_1 fares better than society S_2. On the other hand, if we are a utilitarian, we would have chosen a merit function that measures the total wealth of each society $W_U[S(\mathbf{x})] = \sum x_i$; for the above examples, $W_U[S_1(\mathbf{x})] = 7$ and $W_U[S_2(\mathbf{x})] = 8$: by this criterion, the society S_2 is superior to society S_1.

In mathematics, such a weighting criterion W is called a *functional* and the art of finding the solution $S_\alpha(x)$ that maximizes the given functional is called an optimization problem. All areas of fundamental physics are formulated in the optimization framework and the approach is often called a Lagrangian formulation.

Let us come back to our theory of justice. We have two problems : first we have to choose a weighting functional $W[]$ and second we have to find the best $S_\alpha(\mathbf{x})$ maximizing this functional. What I am going to argue is that even if the veil of ignorance enabled us to choose a particular functional $W[]$, there is not enough information to find a "best" solution.

The problem I am considering is a simple numerical one where all the complexity of human behavior have been simplified into some collection of number. I show is that even in this very simplified framework, there is no unique solution. Therefore, we should not hope for a unique solution in the much more complex domain of human behavior.

We will consider below few simple cases and show that people adhering to the same principle of justice but with different theoretical model of wealth production will choose very different organization of the society.

A. Rawlsian individuals.

Let us consider a society formed of only two individuals, which makes the demonstration particularly easy. Demonstrations for the general cases can be found in the appendix. I suppose first that all individuals behind the veil are Rawlsian and adhere to the Maximin principle. There are two position x_1 and x_2 available in the society, where x_1 designates the wealth of the worst-off and x_2 the wealth of the best-off person, hence, $x_1 \leq x_2$.

In the simplest possible model, in a given time interval Δt, individuals produce goods and services. The surplus P of these goods is then shared between individuals and added to their wealth by a simple rule, a proportion α for the worst off and $(1-\alpha)$ for the best off.

We wish to find the best choice for the sharing rule α satisfying the Maximin principle.

Case 1. Neutralist theory. The theoretical model of the individuals in this case is that the output P during interval Δt of the society, whatever the later distribution between its members, is a constant, proportional to the size of society, productivity per person and duration of time interval and : $P = p\Delta t$. This output is shared between individuals and added to their wealth

$$x_1(t + \Delta t) = x_1(t) + \alpha p \Delta t \quad (1)$$

$$x_2(t + \Delta t) = x_2(t) + (1 - \alpha) p \Delta t \quad (2)$$

where t designates the time, Δt the time interval during which the output P is produced, and $x_i(t)$ is the wealth of individual i at time t.

To find a society $S(\mathbf{x}) = \{x_1, x_2\}$ that maximizes x_1 (in a given time window) is, in mathematical term, finding $\{x_1, x_2\}$ such that

$$W_R[\{x_1, x_2\}] = x_1 \text{ is maximum} \quad (3)$$

$$x_1 \leq x_2 \quad (4)$$

This is an optimization problem (equation 3) with a constraints (equations 4).

The above problem is particularly simple as individuals increase their wealth linearly as a function of time

$$x_1(t) = \alpha p t \ ; \ x_2(t) = (1 - \alpha) p t$$

It is obvious here that the best sharing rule satisfying relations (3,4) is

$$\alpha = \frac{1}{2}$$

i.e. a totally egalitarian society. The demonstration generalizes trivially to N individuals (appendix A).

Case 2. trickle-down theory. The individuals in this category believe that the output P of the society is enhanced by some amount of inequality that spurs competition. Let us consider the very simple model where the total output is the same constant as before, plus a simple measure of inequality :

$$P = [p + \mu(x_2 - x_1)] \Delta t$$

where μ is a coefficient weighting the importance given to inequality as the enhancer of wealth production. As before, this output is shared between society members according to a sharing coefficient α:

$$x_1(t + \Delta t) = x_1(t) + \alpha [p + \mu(x_2 - x_1)] \Delta t \quad (5)$$

$$x_2(t + \Delta t) = x_2(t) + (1 - \alpha) [p + \mu(x_2 - x_1)] \Delta t \quad (6)$$

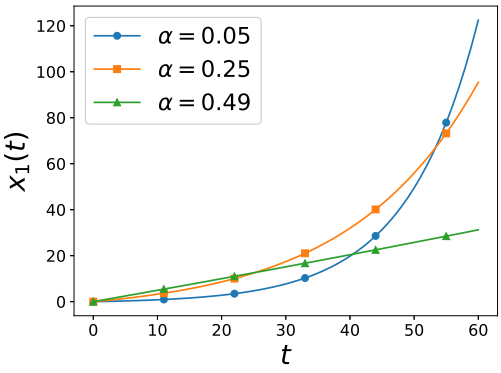

Figure 2. The wealth of the worst-off person for trickle down theory (equations 5-6) grows exponentially if $\mu > 0$:$x_1(t) = p\alpha\left(\exp(\kappa t) - 1\right)/\kappa$ where $\kappa = \mu(1-2\alpha)$ (See appendix). For a fixed value of $p = 1$ and $\mu = 0.1$, three different growth curves corresponding to different inequality values α are shown. We observe that eventually high inequality curves (lower α) take over low inequality ones ($\alpha \approx 1/2$).

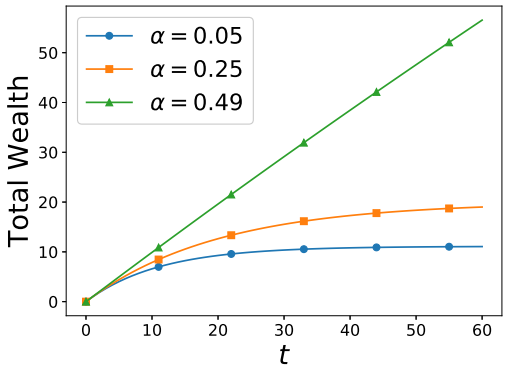

Figure 3. The total wealth of the society, (equations 5-6) is $x_1(t) + x_2(t) = p\left(\exp(\kappa t) - 1\right)/\kappa$ where $\kappa = \mu(1-2\alpha)$ (See appendix) and saturates exponentially to $p/\left|\kappa\right|$ when inequality negatively affects wealth production ($\mu < 0$). For a fixed value of $p = 1$ and $\mu = -0.1$, three different growth curves corresponding to different inequality values α are shown. We observe that more egalitarian rules are better in this case. For a totally egalitarian sharing rule ($\alpha = 1/2$), growth increases linearly without saturation.

As before, we search for a Maximin solution satisfying relation (3-4). However, this time, if we believe that inequality enhances wealth production, *i.e.* $\mu > 0$, from relation (5-6) we deduce that wealth accumulation this time is exponential (figure 2 and appendix). The best choice for α depends on the time window we are considering. For example, if we are considering long term perspective (Keynes adage [9] notwithstanding), a value of $\alpha \approx 0$ has to be chosen : nearly all of the output to the best-off individual. If the Rawlsian individual believes in a theoretical model of wealth production where $\mu > 0$, then he would advocate a society where inequality is *as high as possible*.

B. Utilitarian.

We saw in the above section that a Rawlsian individual will adhere to very different rules of wealth distribution based on his inner theoretical model of wealth production. There is nothing particular to the Maximin principle and a utilitarian will have the same indeterminacy. Let us revisit the above example from the point of view of a utilitarian person that tries to maximize the total wealth. If he is a neutralist (i.e. relation 1-2), he will conclude that the total wealth grows linearly

$$x_1 + x_2 = pt$$

and the rules of output redistribution are of no consequence. On the other hand, if he believes that inequality positively influences wealth production (relation 56 with $\mu > 0$), he will choose an unequal society very similar to a Rawlsian individual with the same inner theoretical model.

However, if the utilitarian believes that inequality negatively affects wealth production, he would choose the most egalitarian society, *i.e.* $\alpha = 1/2$ (figure 3).

IV. CONCLUSION AND DISCUSSION

In mathematics, when n unknown quantities are constrained by m relations and $m < n$, the problem is in general *indeterminate* and has no unique solution[10]. I argued in this article that the "veil of ignorance", as valuable as it is, does not bring enough constraints to determine uniquely the set of rules individuals can agree to. Indeed, even when people agree on a principle of justice (such as Maximin), they can choose diametrically opposing set of rules for the organization of the society. Because Rawls had stated his theory in very precise and falsifiable terms, it was possible to use basic tools of mathematics to demonstrate the above statement.

The above demonstration can be summarized as the following : Individuals behind the veil of ignorance hold different opinions and theoretical models ; the differences in opinions are not removed by the veil and lead to failure in reaching an agreement if the original diversity of opinion is large. This is true even if they accept the same principle of justice. Mathematically, we showed that given a principle of justice $W[]$ which can weight possible societies S_α ($W[S_\alpha]$ is the numerical *merit* of the society S_α), no agreement can be reached on the set of rules $\mathcal{R}$ that would organize the society.

In this article, I restricted the discussion to the (un-)determination of the rules $\mathcal{R}$ given the principle of justice $W[]$. But can an agreement be reached on the principle of justice $W[]$ itself ? The problem is very similar to the problem treated here (it is an optimization problem) but at another abstraction level. An agreement on

$W[]$ can be reached if the individuals have the same theoretical model of human behavior. Humans have a bias in evaluating positive and negative outcome, which has been intensely investigated by social psychologists[11, 12], and they may assess that more in relative than absolute terms[13, 14]. If we suppose that individuals are totally risk averse and give much more weight to negative outcome than to positive one, we would naturally choose the Maximin principle, as Rawls did. If we suppose that individuals are neutral in this respect, then we would choose a principle that maximize the mathematical expectation, *i.e.* the total output, as utilitarians do. But variations in individuals risk assessment, which are widespread and may have partially genetic roots[15] shall prevent these individuals to reach an agreement even in the principle of justice.

Mathematically, we could address both these problems by a meta-optimization method : given the variability in the theoretical models of individuals present behind the veil of ignorance, we can compute a solution that minimizes the total dissatisfaction of individuals resulting from the gap between their desired solution and the chosen solution. But then, the individuals have to reach an agreement about how to weight the dissatisfaction ! It seems that there is no rational solution to the multi-level agreement conundrum.

Appendix A: The optimization problem for arbitrary number of individuals.

The wealth accumulation relation (5-6) can be written as a system of ordinary differential equation (ODE)

$$\frac{dx_1}{dt} = \alpha\left(p + \mu(x_2 - x_1)\right) \tag{A1}$$

$$\frac{dx_2}{dt} = (1-\alpha)\left(p + \mu(x_2 - x_1)\right) \tag{A2}$$

We suppose, without loss of generality, that individuals form a new society with zero initial wealth $x_1(0) = x_2(0) = 0$. From the above relations, we deduce that $dx_1/dx_2 = \alpha/(1-\alpha)$ and therefore, at all time,

$$x_1(t) = \frac{\alpha}{1-\alpha} x_2(t)$$

we can use this relation to uncouple the two differential equations :

$$\frac{dx_1}{dt} = \mu(1-2\alpha)x_1 + \alpha p$$

The above equation is a linear first order and its solution is an exponential :

$$x_1(t) = \frac{\alpha p}{\kappa}\left(e^{\kappa t} - 1\right)$$

where $\kappa = \mu(1-2\alpha)$. For $\alpha = 1/2$ (total equality in sharing) or $\mu = 0$ (absence of inequality enhancing wealth production), the growth becomes linear.

If $\mu > 0$, the growth is (positive) exponential and in the long run, the curve with the highest value of κ (highest inequality) will have the highest value of x_1.

On the other hand, if $\mu < 0$, *i.e.* if the theoretical model of an individuals asserts that inequality *decreases* wealth production, then in the long term, x_1 saturates to $\alpha p/|\kappa|$ and more egalitarian criteria fares better. For a totally egalitarian society, the growth is linear and unlimited and would be chosen even by an utilitarian individuals.

The model trivially generalizes to N individuals

$$\frac{dx_i}{dt} = \alpha_i\left(p + \mu(x_N - x_1)\right)$$

where $\sum_{i=1}^{N} \alpha_i = 1$. The same method as above is used to solve the system of ODE and

$$x_1(t) = \frac{\alpha_1 p}{\kappa}\left(e^{\kappa t} - 1\right)$$

where $\kappa = \mu(\alpha_N - \alpha_1)$. In this model of wealth production where inequality is measured crudely in terms of difference between the best-off and the worst-off person, if $\mu > 0$ and we are considering very long terms, the best solution for Maximin is the "winner takes all" : $\alpha_N \approx 1$, $\alpha_{i\neq N} \approx 0$.

Other measures of inequality can be used in other theoretical models and they would lead to different rules for sharing the total output.

[1] John Rawls. *A Theory of Justice*. Harvard University Press, Cambridge, Mass, new ed edition, March 2005.

[2] John Rawls. *Justice as Fairness – A Restatement*. Harvard University Press, Cambridge, Mass, 2rev ed edition, 2001.

[3] Russ Shafer-Landau. *Ethical Theory: An Anthology*. John Wiley & Sons, Chichester, West Sussex ; Malden, MA, 2nd edition edition, 2012.

[4] John C. Harsanyi. Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory. *American Political Science Review*, 69(02):594–606, June 1975.

[5] Olatunji A. Oyeshile. A Critique of the Maximin Principle in Rawls ' Theory of Justice. *Humanity & Social Sciences Journal*, 3(1):65–69, 2008.

[6] Michal Wiktor Krawczyk. A model of procedural and distributive fairness. *Theory and Decision*, 70(1):111–128, January 2011.

[7] Anthony Laden. Games, Fairness, and Rawls's A Theory of Justice. *Philosophy & Public Affairs*, 20(3):189–222, 1991.

[8] John Von Neumann, Oskar Morgenstern, Harold William

Kuhn, and Ariel Rubinstein. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, N.J. ; Woodstock, 60th anniversary commemorative edition, 2007.

[9] "In the long run, we are all dead", *A Tract on Monetary Reform* Cambridge University Press (1923), Ch. 3, p. 65.

[10] In most cases, it has an infinite number of solutions.

[11] Guido Peeters and Janusz Czapinski. Positive-Negative Asymmetry in Evaluations: The Distinction Between Affective and Informational Negativity Effects. *European Review of Social Psychology*, 1(1):33–60, January 1990.

[12] Michael Siegrist and George Cvetkovich. Better Negative than Positive? Evidence of a Bias for Negative Information about Possible Health Dangers. *Risk Analysis*, 21(1):199–206, February 2001.

[13] Leon Festinger. A Theory of Social Comparison Processes. *Human Relations*, 7(2):117–140, May 1954.

[14] Jerry Suls and Ladd Wheeler. *Handbook of Social Comparison: Theory and Research*. Kluwer Academic/Plenum Publishers, New York, 2000 ed. edition, 2000.

[15] David Cesarini, Christopher T. Dawes, Magnus Johannesson, Paul Lichtenstein, and Björn Wallace. Genetic Variation in Preferences for Giving and Risk Taking. *The Quarterly Journal of Economics*, 124(2):809–842, May 2009.