Is Altruism Futile?
Altruistic Career Choices in General Equilibrium

Martin Mattsson*

March 26, 2026

Preliminary and incomplete. Feedback is highly welcome.

Abstract

When choosing a career, it is common to care not just about income, but also about contributing to a better world. This paper studies the implications of such altruistic preferences in general equilibrium. I propose a model with firms that differ in their negative externality rate and a labor force where some altruistic workers internalize the social cost of the externality. I establish three main results. First, a *futility result*: when altruists are few, their sorting into better firms has no effect on equilibrium outcomes, because general equilibrium substitution—normal workers filling vacancies at high-externality firms—undoes the reallocation. Second, this futility is *fragile*: when firms are homogeneous except in their externality rates, the threshold of altruists needed to make a difference is the employment share of a single firm. This threshold shrinks with the number of firms, even when total employment grows proportionally to the number of firms. A small group of altruists can affect outcomes by *concentrating* at the best employer long before they could affect outcomes by *boycotting* the worst. When altruists exceed the threshold, the equilibrium features assortative matching—altruists sort into the cleanest firms, accept lower wages, and generate a compensating differential schedule across firms—with social welfare increasing in altruism intensity. Third, introducing a non-profit sector funded by altruistic donations *weakens* this assortative matching: individual altruists face a three-way trade-off between earning to give at a high-externality firm, working at a low-externality firm for lower pay, and working at the non-profit, formalizing the "earn to give" logic of the effective altruism movement. I show that in an interior equilibrium, consequentialist altruists are indifferent between all three strategies.

*Department of Economics, National University of Singapore. Email: martin.mattsson@nus.edu.sg. I thank Pete Klenow, Bobby Pakzad-Hurson, and Philip Trammell for helpful discussions. This research was supported financially by the National University of Singapore. This paper was written with the assistance of Claude (Anthropic).

1 Introduction

Suppose a talented young student, motivated by a desire to help others, chooses to pursue a career in medicine rather than a more lucrative one in finance. Most people would regard this as an admirable choice—one that makes the world a better place. But does it? The medical school seat is scarce: by accepting her place, the student displaces another applicant, who takes a job in finance instead. The same number of doctors graduate, the same number of patients are treated, and the same number of people work in finance. The altruistic student has accepted a lower lifetime income without improving aggregate outcomes. Her sacrifice is real, but its social impact is zero: she has merely switched places with someone else.

This "replacement" logic is not confined to medicine. Whenever altruistic workers sort into socially beneficial jobs, the positions they vacate are filled by others. In a competitive labor market with interchangeable workers, general equilibrium substitution can undo the effect of altruistic sorting entirely. The question, then, is not whether individual career choices feel meaningful, but whether they change equilibrium outcomes.

A growing body of evidence documents that workers do, in fact, accept lower wages to work for employers perceived as socially responsible: Burbano (2016) shows that corporate social responsibility messages reduce workers' wage requirements in field experiments, Hedblom et al. (2019) find that advertising a job as socially oriented attracts more and more productive applicants, and Krueger et al. (2024) document a substantial sustainability wage gap in administrative data, with stronger effects for high-skill workers. Non-profit organizations attract workers at significant wage penalties, consistent with "labor donation" by mission-motivated employees (Preston, 1989; Gregg et al., 2011). These wage penalties are substantial and well-documented. But whether they translate into social welfare gains—or are simply private sacrifices with no aggregate effect—is an open question that requires a general equilibrium analysis.

This paper is, to my knowledge, the first to study the labor market implications of altruistic preferences in general equilibrium. I develop a model with firms that produce the same consumption good but differ in the negative externalities they generate. Workers are either normal (ignoring the externality in their choices) or altruistic (partially or fully internalizing the social cost). A key distinction throughout the paper is between *behavioral utility*—the objective function that governs workers' choices—and *experienced utility*—the welfare-relevant payoff common to all workers. I show how the choices of a rational consequentialist utilitarian worker can be described by simply reducing the utility linearly by the total externality.

The paper establishes three main results. First, *futility.* In the simplest version of the

model with just two firms, it does not matter for social welfare if altruistic workers choose to only work for the firm producing the least negative externality, as long as the number of altruists is less than half of the workforce. Such sorting has no effect on equilibrium employment, wages, output, or the externality. General equilibrium substitution—normal workers filling the vacancies at high-externality firms—undoes the reallocation. The intensity of altruism does not affect social welfare.

Second, *reversal of futility with many firms.* The futility result is fragile. When firms are homogeneous except in their externality rates, the threshold for futility is the employment share of a single firm, which shrinks inversely with the number of firms—even when total employment grows proportionally. A small group of altruists can break through futility by *concentrating* at the best employer long before they could affect outcomes by *boycotting* the worst. With 20 firms of equal size, concentration requires just 5% of the workforce while boycotting requires 95%. This asymmetry has implications for how altruistic workers should organize: identifying the best employer (a strategy pursued by organizations like 80,000 Hours; Todd 2023) is far more effective than identifying the worst (as in polluter rankings such as PERI's Toxic 100 and Greenhouse 100 indices; Ash and Boyce 2024). For a small group of motivated workers, knowing which employer is best matters far more than knowing which is worst. When altruists are sufficiently numerous, the equilibrium features *assortative matching*: altruists sort into the cleanest firms, accepting lower wages, while high-externality firms become undercrowded and pay more. Wages form a compensating differential schedule, with cleaner firms paying less and social welfare increasing in the intensity of altruism. In this setting, a simple deontological "do no harm" rule—refusing to work at any firm that generates externalities—produces a very similar outcome to utilitarian altruism, because both approaches lead altruists to sort into the cleanest firms.

Third: *the non-profit weakens sorting.* These results change when the model includes a non-profit sector funded by altruistic donations. The non-profit weakens assortative matching: the sorting condition changes from a power-law in the altruism parameter to linear with slope that depends only weakly on altruism intensity, as the donation channel partially absorbs the sorting motive. Individual altruists face a three-way trade-off—between "earning to give"[1] at a higher-externality firm, working at a cleaner firm for lower pay and giving less, and working at the non-profit—and in interior equilibrium are indifferent between all three strategies. The model formalizes the earn-to-give strategy proposed by some in the effective altruism movement (MacAskill, 2015), providing the first general equilibrium foundation for

[1] "Earning to give" refers to the strategy of choosing a high-paying career and donating a substantial portion of one's income to effective charities, rather than working directly in the social sector. See `https://80000hours.org/articles/earning-to-give/` for an overview.

the claim that altruistic workers' firm-sorting choices may be irrelevant while their donation decisions are not.

In this setting, the actions of consequentialists and deontologists start to diverge. Consequentialists are willing to work at higher-externality firms and donate the wage premium to the non-profit, while deontologists following a "do no harm" rule refuse to make such a trade-off, working only at firms with no externalities or at the non-profit. Because deontologists forgo the higher wages available at high-externality firms, they donate less to the non-profit, resulting in lower overall social welfare—even though the deontologic altruists themselves are engaged in less harmful activity.

1.1 Related Literature

This paper relates to several strands of the literature. A growing body of work studies the market consequences of ethical or socially responsible behavior. On the demand side, Kaufmann et al. (2024) show that even consumers who fully internalize an externality overconsume externality-generating goods in competitive equilibrium, because equilibrium feedback dampens each person's impact on aggregate consumption. Trammell (2025) studies altruism in general equilibrium from the consumer side, showing that the supply impact of ethical consumption nonetheless remains significant even in large economies. On the investment side, Pástor et al. (2021) develop an equilibrium model of sustainable investing in which investors with environmental, social, and governance preferences tilt their portfolios toward green assets, generating a "green premium" that lowers the cost of capital for sustainable firms. Arnold (2023) establishes conditions under which socially responsible investing is neutral: when financial markets are complete and firms maximize shareholder value, investors who screen out "bad" stocks have no effect on real outcomes, because other investors purchase those stocks at unchanged prices. The mechanism is closely analogous to the futility result developed here: in both cases, perfect substitutability between ethical and non-ethical agents—complete financial markets in Arnold's setting, interchangeable workers in mine—allows the non-ethical agents to undo the reallocation, leaving equilibrium outcomes unchanged. In Arnold's framework, neutrality breaks down when financial markets are incomplete, so that non-ethical investors cannot perfectly substitute for the screened assets. The fragility result in the present paper is quite different: futility breaks *even with perfect substitutability among workers*, simply because the threshold for altruists to fill an entire firm shrinks as the number of firms grows. With many firms, a small group of altruists can overwhelm a single employer and affect real outcomes—a channel that has no direct counterpart on the investment side, where market completeness ensures neutrality

regardless of the number of assets or the number of ethical investors. The present paper complements these by studying the labor supply side. The futility result developed here has a conceptual parallel in the replaceability argument of MacAskill (2014), who argues that charity workers are replaceable—someone else would fill the position—while philanthropists are not, since their replacements donate far less. In my model, the mechanism is analogous: altruistic workers who sort into the clean firm are replaced by normal workers at the high-externality firm, leaving equilibrium outcomes unchanged. With few firms, a strong futility result emerges: altruistic workers' firm choices have no effect on outcomes due to factor substitution. With many firms, however, the threshold for futility shrinks dramatically, and the key result is an asymmetry between concentration and boycotting strategies. Unlike on the demand side (Kaufmann et al., 2024), no dampening of altruistic impact arises when labor is supplied inelastically. Therefore, when the entire workforce consists of perfect utilitarians who each place equal weight on every other worker's well-being, the Nash equilibrium achieves the first-best allocation of workers to firms—though not the full first best, which would additionally require transfers to equalize consumption across the resulting wage distribution.

The utilitarian microfoundation developed in Section 2.4 fills a gap in the existing literature on other-regarding preferences. Several of the papers cited above—Kaufmann et al. (2024), Trammell (2025), Bergstrom et al. (1986), and Kotchen (2006)—assume reduced-form preferences over externalities or public goods as primitives, without deriving them from deeper welfare considerations. Andreoni (1988) shows that under pure altruism, the fraction of contributors to a public good shrinks to zero as the economy grows—a result that motivated the introduction of warm-glow preferences as an alternative. The microfoundation here offers a different resolution: starting from a utilitarian agent who weights others' *experienced* utility, I show that the agent's incentive to act on the externality remains finite in large economies—the concern for a large number of others, each harmed by the externality, exactly offsets the dilution of individual impact. This yields the reduced-form behavioral utility function with a precise mapping between the altruism parameter and the utilitarian weight, without requiring warm-glow or deontological assumptions.

The model's donation channel relates to the literature on private provision of public goods. Bergstrom et al. (1986) show that voluntary contributions to a public good are strategic substitutes when contributors have quasi-linear preferences, leading to underprovision and crowding out. Warr (1983) establishes a neutrality result: the equilibrium level of the public good is independent of the income distribution among contributors. Kotchen (2006) studies "green markets" as a mechanism for private provision of public goods, showing that the introduction of a green product can have ambiguous effects on environmental

quality. In my model, the linearity of the externality in altruistic utility eliminates strategic substitutability entirely—each altruist's optimal donation is independent of others' donations (see Remark 9). This is a simplifying feature, but I also show that it is possible to ground this altruistic behavioral utility function in a utilitarian concern for the utility of all other agents in the model as the number of agents grow large.

François (2000) studies how public-service motivation can be better harnessed in government than in profit-maximizing firms, providing an efficiency rationale for low-powered incentives in organizations producing social services. Besley and Ghatak (2005) study competition and incentives with motivated agents, showing that matching between principals and agents on mission preferences increases organizational efficiency. While they focus on productivity benefits of mission-matching, I study the general equilibrium consequences of worker sorting for externalities. Delfgaauw and Dur (2007) and Handy and Katz (1998) analyze how firms screen for intrinsically motivated workers, while Brekke and Nyborg (2008) show that firms can use corporate social responsibility as a labor market screening device. On the empirical side, Fehrler and Kosfeld (2014) and Cassar (2019) provide experimental evidence that pro-social missions attract motivated workers who accept lower pay, and Ashraf et al. (2020) study sorting, selection, and productivity in public-service recruitment, finding that career incentives attract more talented applicants without sacrificing prosocial motivation. Morduch and Szafarz (2019) study the occupational choice of altruistic workers between for-profit employment (earning to give) and nonprofit employment (donating labor), showing that the optimal strategy depends on whether budget constraints bind. The present model abstracts from screening and productivity differences to isolate the pure general equilibrium effect of altruistic sorting.

On prosocial behavior more broadly, Bénabou and Tirole (2006) develop a theory combining altruism with reputational concerns, and Andreoni (1990) introduces the distinction between pure altruism and warm-glow giving. The non-profit channel relates to the warm-glow literature in that altruistic donations have a direct, non-negligible impact on outcomes even in large economies, connecting also to the broader literature on externalities and public goods (Cornes and Sandler, 1996).

2 Model

2.1 Firms and the Non-Profit Sector

There are $J \geq 2$ competitive for-profit firms, each producing the same consumption good using labor as its sole input. Firm j's production function is

$$Y_j = AL_j^{\alpha}, \qquad \alpha \in (0,1), \tag{1}$$

where $A > 0$ is the common productivity parameter and L_j is firm j's labor input. Each firm generates an externality proportional to its labor input, at rate $\phi_j \geq 0$. While the externality could be either positive or negative, I normalize so that the best firm generates no externality, $\phi_{\min} = 0$; I call this firm "the clean firm." The externality is not priced: firms bear no cost for generating it. For concreteness, one may think of the consumption good as electricity and ϕ_j as an air pollution emissions rate, with cleaner firms (e.g. a solar power company) having low or zero ϕ_j and high-externality firms (e.g. a coal power plant) having high ϕ_j.

Firms are competitive and pay workers their marginal product of labor:

$$w_j = \alpha A L_j^{\alpha-1}. \tag{2}$$

Workers' only source of income is their wage, so $c = w_j$ for a worker employed at firm j (before any donations).[2]

The economy also contains a non-profit organization that employs workers to generate a positive externality (e.g., environmental remediation or bed nets to prevent malaria). Define the total harm from for-profit production and the total good from the non-profit as

$$H = \sum_{j=1}^{J} \phi_j L_j, \qquad G = \psi L_{\text{np}}, \tag{3}$$

where L_{np} is non-profit employment and ψ is its productivity in abatement. The net externality is

$$E = H - G = \sum_{j=1}^{J} \phi_j L_j - \psi L_{\text{np}}. \tag{4}$$

The non-profit pays its workers a wage $w_{\text{np}} \leq w^*$. When $w_{\text{np}} < w^*$, normal workers strictly prefer the for-profit sector, so only altruistic workers accept employment at the non-profit.

[2]Profits $\pi_j = (1-\alpha)AL_j^{\alpha}$ accrue to firm owners, who are assumed to be a separate class that does not participate in the labor market. This assumption is inessential for the results.

Altruistic workers' contributions to the non-profit take two forms: working at the non-profit at below-market wages (a "labor donation" of $w^* - w_{\text{np}}$ per worker), and monetary donations from wages earned at for-profit firms. Total non-profit employment is

$$L_{\text{np}} = \frac{D + L^a_{\text{np}}(w^* - w_{\text{np}})}{w^*}, \tag{5}$$

where D is total monetary donations and L^a_{np} is the number of altruists working at the non-profit at a lower wage. As shown in Section 5, these two channels are equivalent in the frictionless model: the equilibrium allocation depends only on total effective contributions, not on their form.

2.2 Workers

There are L workers, each of whom inelastically supplies one unit of labor. A share $M \in (0, 1)$ of workers are altruistic, with a common altruism parameter $\delta > 0$; the remaining share $1-M$ are normal workers with $\delta = 0$.

Two utility concepts must be distinguished. Behavioral utility governs workers' choices. For normal workers, behavioral utility depends only on consumption:

$$u_k = v(c), \tag{6}$$

where $v(c) = c^{1-\sigma}/(1-\sigma)$ with $\sigma > 1$ is the CRRA utility function.[3] For altruistic workers, behavioral utility also depends on the net externality:

$$u_k = v(c) - \delta E. \tag{7}$$

Crucially, altruistic workers care about the *total* externality E, not their individual contribution to it. This makes them *consequentialists*: they evaluate actions by their consequences for the externality, not by whether the action itself satisfies a moral rule. The model does not, however, require that the externality represent harm to other workers in the economy. The parameter E could capture harm to third parties outside the model—animal suffering, the health of individuals in other countries, environmental damage to future generations, or any outcome the altruist deems morally relevant. All results in the main text hold under

[3]The case $\sigma = 1$ corresponds to log utility. With log utility and donation recycling through the non-profit (Section 5), the ratio of the marginal return to donations to the marginal utility of consumption is constant at $\delta\psi$, independent of consumption. For $\delta\psi > 1$, no interior donation equilibrium exists. With $\sigma > 1$, the marginal utility of consumption grows faster than the marginal return to donations as consumption falls, yielding the well-defined interior solution $c^*_a = (\delta\psi)^{-1/(\sigma-1)}$. I set $\sigma = 2$, a standard calibration in the macroeconomics literature.

any interpretation that generates the reduced-form behavioral utility (7).

One natural microfoundation is *utilitarian*: altruists internalize other workers' disutility from the externality (Section 2.4). Under this interpretation, $\delta = \lambda L\gamma$ where $\lambda \in (0, 1]$ is the weight the altruist places on each other worker's utility. The upper bound $\lambda = 1$ ($\delta = L\gamma$) corresponds to a *perfect utilitarian*—a worker who maximizes the sum of all workers' utilities, as in the classical utilitarian tradition (Bentham, 1789; Mill, 1863; Sidgwick, 1874). At this bound, the altruist approximately maximizes social welfare in the large-economy approximation; when all workers are altruistic ($M = 1$), the Nash equilibrium approximately achieves the first best (see Section 2.4).

Experienced utility is the welfare-relevant payoff. Every worker—normal and altruistic alike—is harmed by the externality (e.g., everyone breathes polluted air). Experienced utility is

$$\tilde{u}_k = v(c_k) - \gamma E, \tag{8}$$

where $\gamma > 0$ is the per-worker disutility from the externality. Normal workers' behavioral and experienced utilities differ because, in a large economy, each worker's individual contribution to E is too small to justify costly avoidance, so they ignore the externality in their choices—a tragedy of the commons (see Section 2.4). Altruistic workers' behavioral and experienced utilities differ because $\delta \neq \gamma$ in general. The behavioral parameter δ reflects the altruist's moral concern for the externality's consequences, not additional personal harm.

2.3 Social Welfare

I define social welfare as the sum of all workers' experienced utilities:

$$W = \sum_{k=1}^{L} \tilde{u}_k = \sum_{k=1}^{L} v(c_k) - L\,\gamma\,E. \tag{9}$$

This welfare function is the sum of individual experienced utilities—the classical utilitarian criterion, recently given an axiomatic foundation by Alexandrie (2025)—and is not a separate social planner's objective. As shown in Section 2.4, in the large-economy approximation the upper bound $\delta = L\gamma$ ($\lambda = 1$) corresponds to a perfect utilitarian who approximately maximizes social welfare; when all workers are altruistic ($M = 1$), the Nash equilibrium approximately achieves the first-best allocation of workers to firms. However, it does not achieve the full first best, because the market generates unequal wages across firms (compensating differentials), and with concave utility this consumption inequality is welfare-reducing. The true first best would combine the same allocation with lump-sum transfers to equalize consumption (see Remark 6).

2.4 Utilitarian Microfoundation

This section provides a *utilitarian* microfoundation for the behavioral utility functions above. I show that: (i) in a large economy, normal workers approximately maximize $v(c)$—a tragedy of the commons; and (ii) altruistic workers' effective utility is approximately $v(c) - \delta E$ with $\delta = \lambda\Gamma$, where $\lambda \in (0, 1]$ is the weight the altruist places on each other worker's utility and $\Gamma = L\gamma$ is the aggregate social harm. At the upper bound $\lambda = 1$, the altruist is a perfect utilitarian: an agent who seeks to maximize the sum of all individuals' utilities, weighting each person's welfare equally—or in Sidgwick's formulation, "the good of any one individual is of no more importance ... than the good of any other."

As emphasized above, this utilitarian microfoundation is not the only way to rationalize the behavioral utility functions. The reduced-form $v(c) - \delta E$ also arises if the externality represents harm to third parties outside the model. The main text results hold under any interpretation that generates the same reduced-form preferences.

2.4.1 Normal Workers: Tragedy of the Commons

Each worker's experienced utility is $\tilde{u}_j = v(c_j) - \gamma E$ where $\gamma = \Gamma/L$, with $\Gamma > 0$ the aggregate social harm parameter.[4]

When worker j switches from the good firm to the bad firm, the externality increases by ϕ. The private cost to worker j through the externality term is $\Delta\tilde{u}_j^{\text{ext}} = -(\Gamma/L) \cdot \phi$. In a large economy, this per-person harm from one additional unit of externality-generating labor is negligible.

Lemma 1 (Normal Workers Ignore the Externality)**.** *With per-person externality harm $\gamma = \Gamma/L$, the private cost of switching from a clean to a high-externality firm is $\Gamma\phi/L$. For L large, this is negligible relative to any wage differential, so normal workers' equilibrium behavior is approximately the same as if they maximized $v(c_j)$ alone.*

Proof. In a Nash equilibrium, the marginal normal worker is indifferent between firms. Worker j at firm g compares staying versus deviating to firm b, accounting for the externality change $\Delta E = \phi$: $v(w_b') - v(w_g) = (\Gamma/L)\phi$. For L large, the right-hand side $\Gamma\phi/L$ is negligible, so the equilibrium condition is approximately wage equalization. □

This is a tragedy of the commons: each worker's private cost from the externality ($\Gamma\phi/L$) is negligible for large L, even though the total social cost ($\Gamma\phi$) is finite. Normal workers

[4]The scaling $\gamma = \Gamma/L$ ensures that total social harm from the externality, $L\gamma E = \Gamma E$, remains well-defined regardless of L. This is natural for externalities like air pollution, where the total social cost of a given pollution level is finite and independent of how many people share the economy.

rationally ignore the externality because their individual contribution to it has a negligible effect on their own welfare. This is the same logic that leads Andreoni (1988) to conclude that the fraction of contributors to a public good shrinks to zero as the economy grows. Importantly, this does not depend on general equilibrium adjustment—it holds even fixing all other workers' actions. This distinguishes the tragedy of the commons from the futility theorem (Proposition 2), which relies on general equilibrium substitution.

2.4.2 Altruistic Workers: Internalizing the Social Cost

An altruistic worker j cares not only about their own utility but also about the utility of others:

$$u_j^a = \tilde{u}_j + \lambda \sum_{k \neq j} \tilde{u}_k = v(c_j) - \frac{\Gamma}{L} E + \lambda \sum_{k \neq j} \left[v(c_k) - \frac{\Gamma}{L} E \right], \tag{10}$$

where $\lambda \in (0, 1]$ is the altruism parameter. At $\lambda = 1$, the altruist is a perfect utilitarian who treats all workers symmetrically. Note that the behavioral utility $v(c) - \delta E$ used in the main analysis is a large-economy approximation of this exact objective (10): it retains the own-consumption and externality terms but drops the consumption externality terms $\lambda \sum_{k \neq j} v(c_k)$, which do not affect firm-choice incentives in the limit. This approximation performs well for moderate values of λ but can overstate optimal donations when λ is close to 1, because the exact objective includes the positive spillover of other workers' consumption on the altruist's utility (see Remark 1).

When altruist j switches from the good to the bad firm ($\Delta E = \phi$), the cost through the externality terms is:

$$\Delta u_j^{a,\text{ext}} = -\frac{\Gamma}{L} \phi \left[1 + \lambda (L - 1) \right] \approx -\lambda \Gamma \phi \quad \text{for large } L. \tag{11}$$

Defining $\delta \equiv \lambda \Gamma$, the altruistic worker's switching cost is approximately $-\delta\phi$, which is finite and independent of L. The altruistic worker's concern for L others, each harmed by $\Gamma\phi/L$, produces a total cost of $\lambda\Gamma\phi$—exactly offsetting the dilution of individual impact. The altruistic worker's effective utility over firm choice and donations is therefore approximately $u_a \approx v(c_j) - \delta E + \text{constant}$, which is the behavioral utility (7) adopted in the main analysis.

Remark 1 (Perfect Utilitarianism)**.** *Since $\lambda \in (0, 1]$, the behavioral parameter satisfies $\delta = \lambda\Gamma \in (0, \Gamma]$. At $\lambda = 1$, the altruist's objective* (10) *becomes $u_j^a = \sum_{k=1}^{L} \tilde{u}_k = W$, so the altruist maximizes social welfare. This is the defining commitment of classical utilitarianism. The maximally altruistic worker's optimal donation in the non-profit model coincides with the social planner's optimum (see Section 5.5).*

Roadmap. I analyze this model progressively. Section 3 develops the first result (futility) in the simplest case of two firms. Section 4 develops the second result (reversal of futility), extending to many firms where the concentration-versus-boycotting asymmetry, assortative matching, and the wage ladder emerge. Section 5 develops the third result, introducing the non-profit sector and the three-way trade-off.

3 When Is Altruism Futile?

I begin with the simplest case: two firms ($J = 2$, labeled g and b for "good" and "bad"). The good firm is clean ($\phi_g = 0$) and the bad firm generates externality $E = \phi L_b$ with $\phi > 0$. I work throughout with a finite economy in which each worker's firm choice discretely affects the externality, and characterize the Nash equilibrium of the resulting game. The non-profit sector is shut down ($D = 0$, $L_{\text{np}} = 0$) until Section 5.

3.1 No Altruism Equilibrium

When no worker internalizes the externality—as when all workers are non-altruistic—wages equalize across firms. This pins down a *no altruism equilibrium* that depends only on technology.

Lemma 2 (Wage Equalization)**.** *If wages differ across firms ($w_g \neq w_b$), all normal workers choose the higher-wage firm. Hence, in any equilibrium with both firms operating and $M < 1$, wages must equalize: $w_g = w_b$.*

The wage equalization condition pins down the no altruism equilibrium. Since both firms share common productivity A, setting $\alpha A L_g^{\alpha-1} = \alpha A L_b^{\alpha-1}$ gives equal employment:

$$L_g^* = L_b^* = \frac{L}{2}, \tag{12}$$

with common equilibrium wage

$$w^* = \alpha A \left(\frac{L}{2}\right)^{\alpha-1} \tag{13}$$

and equilibrium externality

$$E^* = \phi L_b^* = \frac{\phi L}{2}. \tag{14}$$

Remark 2. *The no altruism allocation* (12) *depends only on total labor supply L and the technology parameters (A, α). It does not depend on the altruist share M or the altruism parameter δ. I use the notation (L_g^*, L_b^*, w^*, E^*) to denote this technology-determined allocation.*

3.2 Nash Equilibrium with Altruistic Workers

I now characterize the Nash equilibrium of the firm-choice game. Each worker simultaneously chooses a firm to maximize their utility, taking other workers' choices as given. In a finite economy, each worker's choice discretely affects the externality by $\pm\phi$.

Proposition 1 (Nash Equilibrium Characterization)**.** *In the unique stable Nash equilibrium:*

1. ***When*** $M \leq 1/2$***:*** *All altruists work at the good firm. Normal workers fill both firms to equalize wages. The equilibrium allocation* (L_g^*, L_b^*)*, wage* w^**, and externality* $E^* = \phi L/2$ *are identical to the no altruism equilibrium.*

2. ***When*** $M > 1/2$***:*** *Let* $M_g(\delta)$ *denote the number of altruistic workers at the good firm. Then* M_g *solves the marginal altruist's indifference condition*

$$\delta\phi = v(w_b) - v(w_g), \tag{15}$$

where $w_g = \alpha A M_g^{\alpha-1}$ *and* $w_b = \alpha A(L - M_g)^{\alpha-1}$*. The remaining* $(1 - M)L$ *normal workers and* $ML - M_g$ *altruists are at the bad firm. The externality is* $E = \phi(L - M_g)$*.*

With CRRA utility, the interior equilibrium $M_g(\delta)$ *is characterized by* (15) *and must be computed numerically.*

Full coordination ($M_g = ML$*, all altruists at the good firm) obtains if and only if*

$$\delta\phi \geq v\Big(\alpha A\big((1 - M)L\big)^{\alpha-1}\Big) - v\big(\alpha A(ML)^{\alpha-1}\big)\,. \tag{16}$$

Proof. Case 1: $M \leq 1/2$. Normal workers equalize wages across firms (Lemma 2), producing the no altruism allocation $L_g^* = L_b^* = L/2$. Since $ML \leq L/2$, all altruists fit at the good firm alongside $L/2 - ML$ normal workers. In this candidate equilibrium, wages are equalized ($w_g = w_b = w^*$) because normal workers are present at both firms. Consider any altruist: switching from the bad firm to the good firm yields a behavioral gain of $\delta\phi > 0$ (reducing E by ϕ) with a consumption change of only $O(1/L)$ (from the discrete reallocation of one worker). So every altruist strictly prefers the good firm. With all altruists at the good firm, wages remain equalized and the allocation is identical to the no altruism equilibrium.

Case 2: $M > 1/2$. When $M_g > L/2$ altruists are at the good firm, the good firm is overcrowded: $w_g < w_b$. Normal workers strictly prefer the bad firm. An altruist at the good firm contemplating a switch to the bad firm gains $v(w_b) - v(w_g) > 0$ in consumption utility but incurs a behavioral cost $\delta\phi$ from increasing the externality. The deviation is unprofitable when $\delta\phi \geq v(w_b) - v(w_g)$. An altruist at the bad firm contemplating a switch to the good

firm gains $\delta\phi$ but loses $v(w_b) - v(w_g)$. Both conditions are satisfied simultaneously at M_g where $\delta\phi = v(w_b) - v(w_g)$.

With CRRA utility $v(c) = c^{1-\sigma}/(1-\sigma)$ and common A, the indifference condition (15) determines M_g implicitly; the solution must be computed numerically.

Uniqueness and stability. As M_g increases from L_g^*, w_g falls and w_b rises, so $v(w_b) - v(w_g)$ is strictly increasing in M_g. This means the indifference condition $\delta\phi = v(w_b) - v(w_g)$ has a unique solution, and the equilibrium is stable: if slightly too many altruists are at the good firm, the wage gap exceeds $\delta\phi$ and some switch back. □

Remark 3. *At $\delta = 0$ (no altruism), the indifference condition gives $M_g(0) = L_g^*$: the interior equilibrium reproduces the no altruism allocation. As δ increases, $M_g(\delta)$ increases continuously, and the externality $E = \phi(L - M_g)$ falls smoothly from E^* toward $\phi(1-M)L$ (at full coordination). There is no discrete jump between "no altruism" and "coordinated" equilibria—the Nash equilibrium traces a smooth path parameterized by δ.*

3.3 The Futility of Altruism

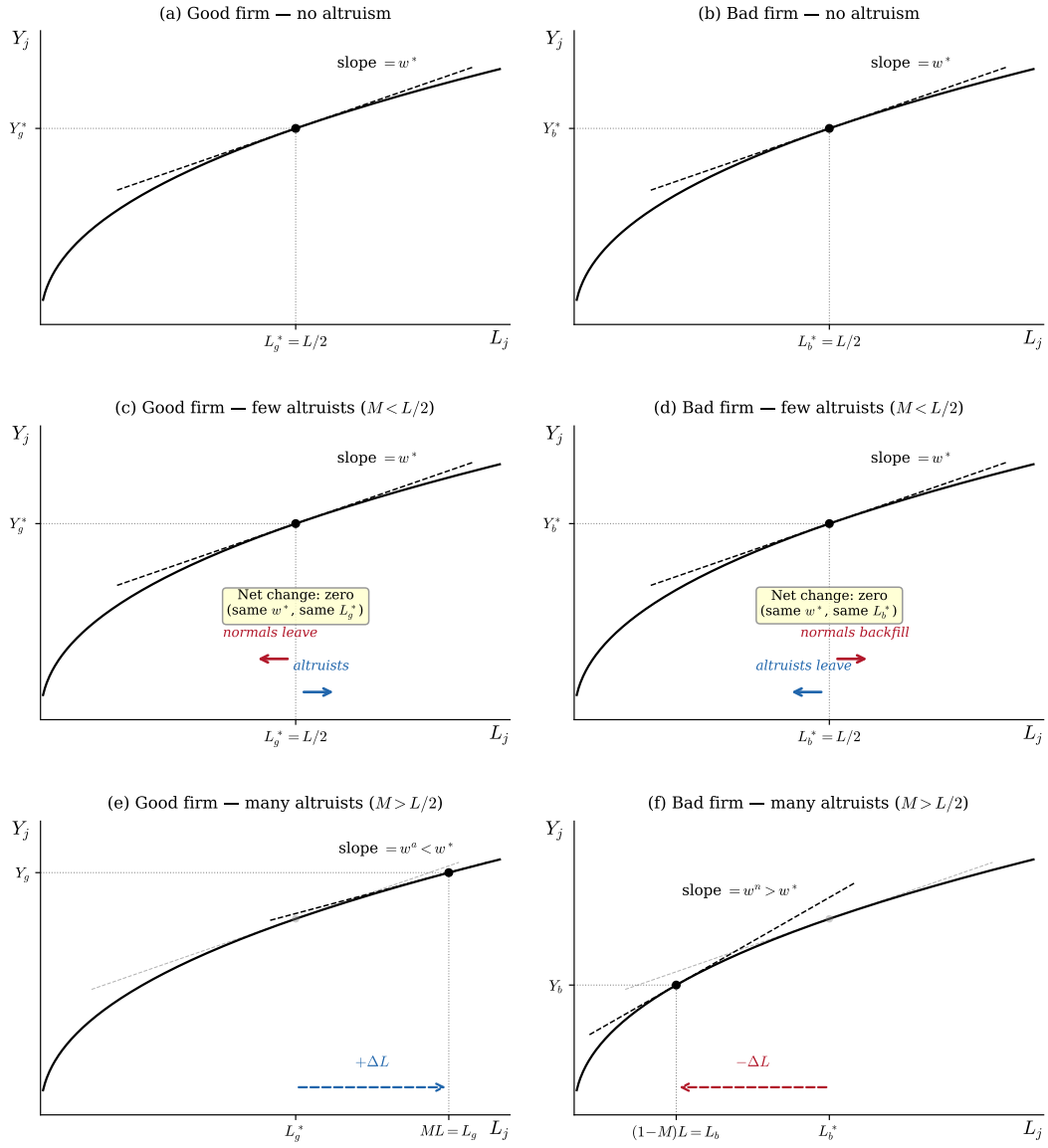
Proposition 1 establishes that altruistic workers optimally sort into the good firm in any Nash equilibrium. When $M \leq 1/2$, however, this sorting has no effect on outcomes—a stronger statement than merely assuming altruists are randomly allocated.

Proposition 2 (Futility of Altruism)**.** *In the model without a non-profit, when $M \leq 1/2$:*

1. *The equilibrium wage w^*, employment levels (L_g^*, L_b^*), and externality E^* are all independent of δ and of the altruist share M.*
2. *Social welfare $W = L\,[v(w^*) - \gamma E^*]$ is independent of δ and M.*
3. *Altruistic workers'* behavioral *utility $v(w^*) - \delta E^*$ is decreasing in δ. However, their* experienced *utility $v(w^*) - \gamma E^*$ is independent of δ and identical to normal workers' experienced utility.*

Proof. By Proposition 1, when $M \leq 1/2$, the Nash equilibrium allocation is $L_g^* = L_b^* = L/2$ with common wage w^*—identical to the no altruism equilibrium. These depend only on (L, A, α), not on δ or M. Parts (1) and (2) follow. For part (3), behavioral utility satisfies $\partial[v(w^*) - \delta E^*]/\partial\delta = -E^* < 0$, while experienced utility $v(w^*) - \gamma E^*$ contains no δ and is therefore constant. □

Figure 1: The futility of altruistic sorting and when it breaks down



Notes: Each panel shows the production function $Y_j = AL_j^\alpha$ for one firm, with the wage equal to the slope of the tangent (the marginal product of labor). Top row (a, b): the no-altruism equilibrium, where both firms share employment equally ($L_g^* = L_b^* = L/2$) and pay the same wage w^*. Middle row (c, d): with few altruists ($M < L/2$). Altruists sort into the good firm (blue arrow), but normal workers leave for the bad firm (red arrow) to equalize wages—the net change is zero. This is the futility result: general equilibrium substitution exactly undoes the altruists' sorting. Bottom row (e, f): with many altruists ($M > L/2$). The good firm is overcrowded ($w^a < w^*$) and the bad firm is undercrowded ($w^n > w^*$). Normal workers cannot fully undo the sorting because there are too few of them, so the equilibrium genuinely changes.

Remark 4. *Figure 1 illustrates the futility result graphically. The futility result is striking because altruists are doing exactly what one would want them to do: they sort into the cleaner firm. Yet this sorting has no effect on outcomes, because normal workers fill the vacancies at the bad firm. Altruists' actual experienced welfare $v(w^*) - \gamma E^*$ is identical to that of normal workers. However, the futility result depends on the two-firm structure. As shown in Section 4, when firms are homogeneous except in their externality rates, the threshold for futility is the employment share of a single firm, which shrinks as $1/J$. Even a small group of altruists can break through futility by concentrating at the best employer.*

3.4 When Altruism Matters: The Interior Equilibrium

When $M > 1/2$, the Nash equilibrium features an interior allocation: enough altruists sort into the good firm to create a wage gap that exactly offsets the behavioral benefit of cleaner production. The degree of altruistic sorting—and hence the externality—depends smoothly on δ.

Proposition 3 (Properties of the Interior Equilibrium)**.** *When $M > 1/2$:*

1. *$M_g(\delta)$ is strictly increasing in δ. As δ increases, more altruists stay at the good firm despite the wage penalty.*
2. *The externality $E(\delta) = \phi(L - M_g(\delta))$ is strictly decreasing in δ.*
3. *Full coordination ($M_g = ML$) obtains for $\delta \geq \delta^*$, where*

$$\delta^* = \frac{v\Big(\alpha A\big((1-M)L\big)^{\alpha-1}\Big) - v(\alpha A(ML)^{\alpha-1})}{\phi}. \tag{17}$$

Proof. For part (1), differentiating the indifference condition (15) with respect to δ (using the implicit function theorem) gives $\partial M_g/\partial\delta > 0$ because the wage gap $v(w_b) - v(w_g)$ is strictly increasing in M_g (more altruists at the good firm widens the gap). Part (2) follows from $E = \phi(L - M_g)$ and M_g increasing. Part (3) follows from evaluating (15) at $M_g = ML$. □

Corollary 1. *Altruistic sorting reduces the externality if and only if $M > 1/2$—that is, altruism matters only if altruists are a majority of the workforce.*

Remark 5 (No Dampening of Altruistic Impact)**.** *On the demand side, Kaufmann et al. (2024) show that even consumers who fully internalize an externality overconsume the externality-generating good, because equilibrium price feedback dampens each person's impact*

on aggregate consumption. No analogous dampening arises on the labor supply side when labor is supplied inelastically. With inelastic supply, total employment is fixed at L regardless of wages: workers' firm choices affect the allocation across firms but not the total amount of labor supplied. Consequently, in the approximate model, when all workers are perfect utilitarians ($M = 1$, $\delta = \Gamma$), each altruist maximizes social welfare. With CRRA utility ($\sigma \neq 1$), the Nash equilibrium under perfect utilitarianism no longer achieves the first best in the large-economy approximation, because the planner's additional term $(\alpha-1)w_j v'(w_j) = (\alpha-1)w_j^{1-\sigma}$ varies across firms.[5] *There is no intensive margin through which equilibrium feedback can erode the altruist's impact: each worker contributes exactly one unit of labor to the externality, regardless of the wage. If labor supply were elastic, dampening could re-emerge, as higher wages at the clean firm would attract additional labor through the extensive margin, partially offsetting the altruistic sorting.*[6]

Define $\bar{M}(\delta)$ as the maximum altruist share for which full coordination is sustainable at altruism intensity δ:

$$\bar{M}(\delta) = \max\left\{M : \delta\phi \geq v\Big(\alpha A\big((1-M)L\big)^{\alpha-1}\Big) - v\big(\alpha A(ML)^{\alpha-1}\big)\right\}. \tag{18}$$

The function $\bar{M}(\delta)$ is increasing in δ: more altruistic workers tolerate a larger wage sacrifice to sustain full coordination.

Proposition 4 (Welfare with Altruistic Sorting)**.** *When $M > 1/2$:*

1. *Social welfare is $W = M_g\, v(w_g) + (L - M_g)\, v(w_b) - L\,\gamma\, E$, where $M_g = M_g(\delta)$ is the Nash equilibrium number of altruists at the good firm.*

2. *Both the altruist share M and the intensity of altruism δ affect welfare: M determines the potential for externality reduction; δ determines how much of that potential is realized through sorting.*

[5]The social planner's FOC for allocating labor across firms is $v(w_j) + (\alpha - 1)w_j v'(w_j) - \Gamma\phi_j = \text{const}$. With CRRA utility $v(c) = c^{1-\sigma}/(1-\sigma)$, the second term is $(\alpha - 1)w_j^{1-\sigma}$, which varies across firms when wages differ. The Nash condition—equalizing $v(w_j) - \Gamma\phi_j$ across firms—therefore does not coincide with the planner's. However, in the exact finite model where each altruist's objective is W (Remark 1), the game is a potential game with potential W, and the Nash equilibrium maximizes W over worker allocations. This achieves the best possible allocation of workers to firms, but still falls short of the full first best, which would additionally use lump-sum transfers to equalize consumption across the wage distribution (Remark 6).

[6]With elastic labor supply, the total externality depends on aggregate employment, which responds to equilibrium wages. A shift of altruists toward the clean firm lowers the clean wage (diminishing marginal product) and raises the high-externality wage, making the high-externality firm more attractive to marginal workers along the supply curve. This general equilibrium feedback partially undoes the intended reallocation, creating the same qualitative dampening that Kaufmann et al. (2024) identify on the demand side.

3. *The externality falls smoothly from E^* (at $\delta \to 0$) to $\phi(1-M)L$ (at full coordination), with a correspondingly smooth change in welfare.*

Proof. Part (1) follows from substituting the Nash equilibrium allocation into (9). For part (2), M enters directly (bounding $M_g \leq ML$) and δ determines $M_g(\delta)$ via the indifference condition (15). For part (3), $M_g(0) = L_g^*$ gives $E = E^*$, while $M_g(\delta^*) = ML$ gives $E = \phi(1-M)L$. □

Remark 6 (Sorting Is Not the First Best)**.** *The sorting equilibrium reduces the externality but cannot achieve the first-best allocation. A social planner would redirect labor toward the clean firm* and *use lump-sum transfers to equalize consumption across workers. Without transfers, sorting can only reduce the externality by creating wage inequality: altruists crowd into the good firm, depressing w_g below w_b. Under concave utility, this wage dispersion is socially costly. As M grows large (with δ high enough for full coordination), nearly all workers are crammed into the good firm at very low wages while the few normal workers at the bad firm earn high wages. The marginal externality reduction from one additional altruist at the good firm becomes negligible (the externality is already near zero), but the marginal consumption loss from overcrowding is steep—especially under CRRA with $\sigma > 1$, where $v'(c) = c^{-\sigma}$ grows rapidly as c falls. Consequently, social welfare can* decrease *in M at high altruist shares: the cost of misallocation exceeds the benefit of externality reduction. This non-monotonicity is visible in Figure 2(b) for $\delta = \Gamma$.*

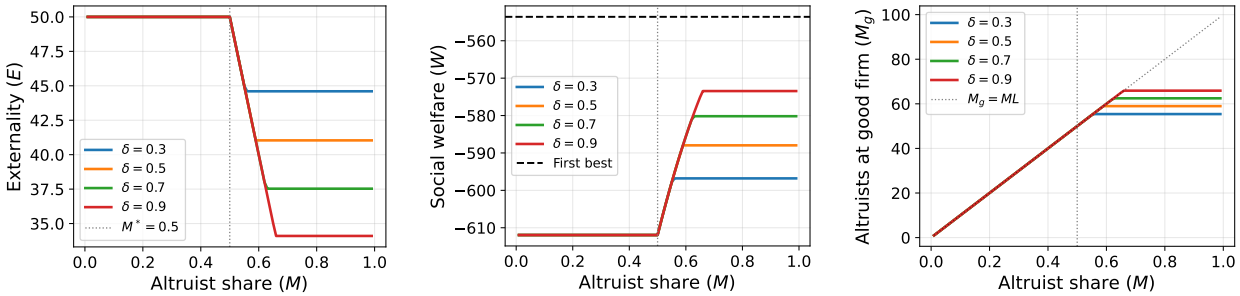
Figure 2 illustrates the Nash equilibrium as the altruist share M varies, for different values of δ. Panel (a) shows that for $M \leq 1/2$, the externality equals E^* regardless of δ (the futility regime). For $M > 1/2$, higher δ leads to more sorting and a lower externality. Panel (b) shows that welfare is non-monotonic in M at high δ: it rises as sorting reduces the externality, peaks, and then declines as the consumption cost of extreme overcrowding at the good firm dominates (Remark 6). Panel (c) shows M_g: for low δ, some altruists remain at the bad firm even when M is large.

Figure 3 shows the interior equilibrium as δ varies, for different values of M. Panel (a) shows $M_g(\delta)$: for $M = 0.20 < 1/2$, $M_g = ML$ for all δ (futility). For $M > 1/2$, M_g increases smoothly with δ until reaching ML (full coordination).

The two-firm model provides the essential building blocks: futility when altruists are few, and an interior equilibrium when they are many. The next section shows how these results change—and become far richer—with many firms.

Figure 2: Nash equilibrium outcomes as M varies

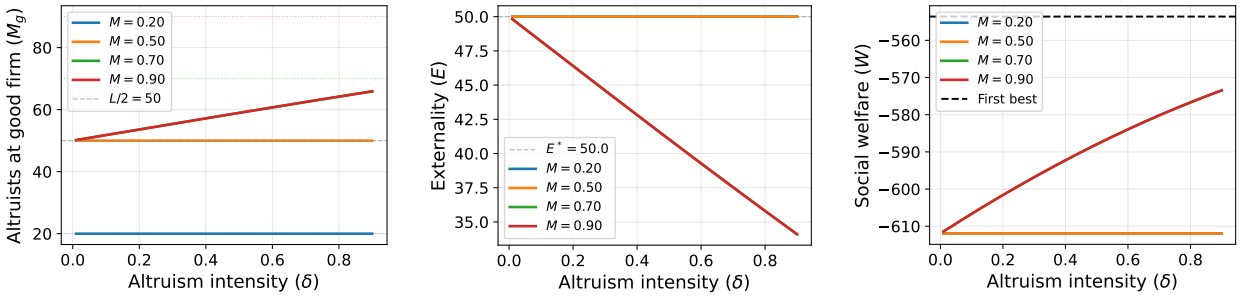
(a) Externality (b) Social welfare (c) Altruists at good firm (M_g)



Notes: For $M \leq 1/2$, all curves overlap: futility holds regardless of δ. For $M > 1/2$, higher δ leads to more sorting (M_g closer to ML), lower externality, and higher welfare. The dotted line in panel (c) shows $M_g = ML$ (full coordination). δ ranges from 0 to $0.30\Gamma = 0.9$ (Appendix D). Parameters: $A = 1.0$, $\alpha = 0.7$, $\phi = 1.0$, $\gamma = 0.03$, $L = 100$, $\sigma = 2$.

Figure 3: The interior equilibrium

(a) Altruists at good firm $M_g(\delta)$ (b) Externality (c) Social welfare



Notes: For $M = 0.20 < 1/2$, outcomes are flat (futility). For $M > 1/2$, the externality falls smoothly and welfare rises as δ increases. δ ranges from 0 to $0.30\Gamma = 0.9$ (Appendix D). Parameters: $A = 1.0$, $\alpha = 0.7$, $\phi = 1.0$, $\gamma = 0.03$, $L = 100$, $\sigma = 2$.

4 The Fragility of Futility

This section develops the second main result: the reversal of futility with many firms. I extend the model to J firms with common productivity A but different externality rates ϕ_j, keeping the same binary worker types as in Section 3. The two-firm futility result extends, but the threshold shrinks dramatically with the number of firms, generating a stark asymmetry between concentration and boycotting. When altruists exceed the threshold, the equilibrium features assortative matching, a compensating wage ladder, and welfare gains from more intense altruism.

4.1 Equilibrium with Many Firms

Firm $j \in \{1, \ldots, J\}$ has production function $Y_j = AL_j^{\alpha}$ and generates externality $E_j = \phi_j L_j$. I normalize the cleanest firm to have zero externality: $\phi_{(1)} = 0$, consistent with the two-firm model where $\phi_g = 0$. Absent a non-profit sector, the net externality is $E = \sum_{j=1}^{J} \phi_j L_j$. Total labor: $\sum_j L_j = L$.

With common productivity A, wage equalization across all firms gives $\alpha A L_j^{\alpha-1} = w$ for all j, so all firms have equal employment:

$$L_j^* = \frac{L}{J}. \tag{19}$$

The common wage is

$$w^* = \alpha A \left(\frac{L}{J}\right)^{\alpha-1}, \tag{20}$$

and the total externality is

$$E^* = \frac{L}{J} \sum_{j=1}^{J} \phi_j = L\,\bar{\phi}, \tag{21}$$

where $\bar{\phi} = (1/J)\sum_j \phi_j$ is the average externality rate.

Proposition 5 (Futility with Many Firms)**.** *With J firms of common productivity A, let $\phi_{\min} = \min_j \phi_j$ and define $M^* \equiv 1/J$. When $M \leq M^*$, the equilibrium allocation $(L_1^*, \ldots, L_J^*)$, wages w^*, and externality E^* are identical to the no altruism equilibrium.*

Proof. When $M \leq 1/J$, all ML altruists are at the cleanest firm in any Nash equilibrium. To see this, consider an altruist at firm k with $\phi_k > \phi_{\min}$. When $ML < L/J$, normal workers are present at every firm including the cleanest, so wages are equalized. Switching to the firm with $\phi_{\min}$ yields a behavioral gain of $\delta(\phi_k - \phi_{\min}) > 0$ with no wage loss—a profitable deviation. When $ML = L/J$, the cleanest firm is filled entirely by altruists with

employment L/J, while normal workers distribute equally across the remaining $J-1$ firms with L/J each; wages are still equalized across all firms. An altruist at a dirtier firm would gain $\delta(\phi_k - \phi_{\min}) > 0$ by switching to the cleanest firm with no wage loss, so this is again a profitable deviation. Hence all altruists must be at the cleanest firm. Since $ML \leq L/J$, they fit at the cleanest firm, wages are equalized, and equilibrium employment at each firm is L/J. □

4.2 Concentration versus Boycotting

With J firms and $M \leq M^*$, all altruists work at the cleanest firm in Nash equilibrium (Proposition 5). With common productivity A, the futility threshold is simply

$$M^* = \frac{1}{J}, \tag{22}$$

the employment share of any single firm. The threshold *decreases* with the number of firms.

This generates a stark asymmetry between *concentration* (altruists sorting into the best firm) and *boycotting* (altruists avoiding the worst firm). To break futility through concentration, altruists need $M > 1/J$. To affect the allocation of the worst firm through boycotting, altruists need $M > (J-1)/J$—enough to leave the worst firm short of workers even after all normal workers have sorted there.[7] Once futility is broken, the per-worker externality reduction is the same for either approach: each altruist who moves from a higher-externality to a lower-externality firm reduces E by $\phi_k - \phi_j > 0$. But the threshold for *any* effect at all is dramatically lower for concentration.

This result has implications for how altruistic workers should organize. Organizations like 80,000 Hours (Todd, 2023) identify the best career paths and encourage workers to seek out high-impact employers—a concentration strategy. By contrast, initiatives such as PERI's Toxic 100 and Greenhouse 100 indices (Ash and Boyce, 2024) identify the worst corporate polluters, providing information that supports a boycotting strategy. The model predicts that concentration is far more effective when altruists are a small minority: with 20 firms of equal size, a group representing just 5% of the workforce can break through futility by concentrating at the cleanest firm, while boycotting the worst would require 95%. For a small group of motivated workers, knowing which employer is best matters far more than knowing which is worst.

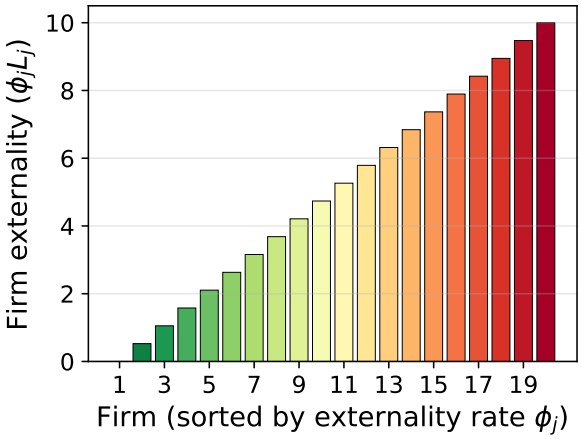
When $M > 1/J$, altruists can no longer all fit at the cleanest firm alongside enough normal workers to maintain wage equalization. The equilibrium features altruistic sorting, analyzed in Section 4.3.

[7] For $J = 20$, concentration requires 5% of the workforce while boycotting requires 95%.

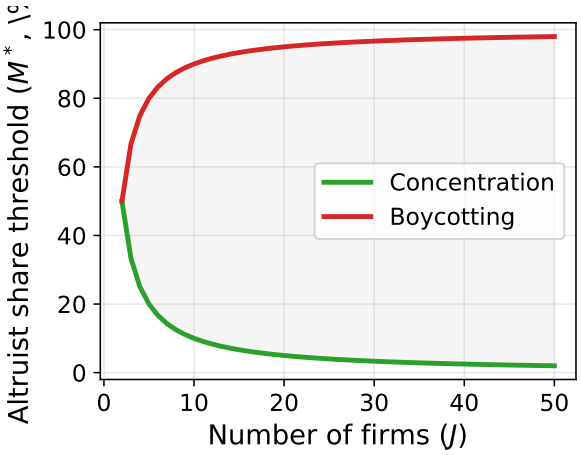
Figure 4 illustrates the many-firms no altruism equilibrium with 20 firms.

Figure 4: Many-firms model

(a) Firm-level externality



(b) Concentration vs. boycott threshold



Notes: Panel (a) shows the no altruism equilibrium with $J = 20$ firms of common productivity A. With equal productivity, all firms have equal employment L/J; firm-level externality varies only through ϕ_j. Firms sorted by ϕ_j; bar color indicates ϕ_j (green = low, red = high). Panel (b) shows the futility threshold for concentration (sorting into the best firm) versus boycotting (avoiding the worst firm), as a share of the workforce, for varying J with L/J held constant. Parameters: $A = 1.0$, $\alpha = 0.7$, $\sigma = 2$.

4.3 Altruistic Sorting

When $M > 1/J$, altruists can no longer all fit at the cleanest firm alongside enough normal workers to maintain wage equalization. The equilibrium features *altruistic sorting*: altruists concentrate at the cleanest firms, accepting lower wages in exchange for a lower externality.

Order firms by externality rate: $\phi_{(1)} < \phi_{(2)} < \cdots < \phi_{(J)}$. In equilibrium, there exists a marginal firm m such that firms $1, \ldots, m-1$ employ only altruists, firm m employs both types, and firms $m+1, \ldots, J$ employ only normal workers.

Proposition 6 (Altruistic Sorting)**.** *In a finite economy with J firms of common productivity A and a share $M > 1/J$ of altruists:*

1. *Normal workers equalize wages across firms $m, m+1, \ldots, J$, giving a common normal-firm wage w_n.*

2. *The altruist indifference condition $v(w_j) - \delta\phi_{(j)} = v(w_n) - \delta\phi_{(m)}$ gives wages at altruist-only firms:*

$$w_j = \left[w_n^{1-\sigma} + (1-\sigma)\delta(\phi_{(m)} - \phi_{(j)})\right]^{1/(1-\sigma)}, \qquad j = 1, \ldots, m-1. \tag{23}$$

3. *Since $\sigma > 1$ and $\phi_{(m)} > \phi_{(j)}$ for $j < m$, this implies $w_j < w_n$: wages at clean firms are strictly below w_n, and employment at clean firms exceeds L/J (altruists are overcrowded, driving down the marginal product).*

4. *High-externality firms are undercrowded ($L_j < L/J$) and pay higher wages ($w_j > w^*$) relative to the no altruism equilibrium.*

5. *The wage gap grows with δ: more altruistic workers accept larger wage penalties to avoid the externality.*

Proof sketch. The $(1 - M)L$ normal workers choose firms to equalize $v(w_j)$: with identical production functions, this requires equal employment and wages at all firms where normals are present. The ML altruists (all with parameter δ) choose firms to equalize $v(w_j) - \delta\phi_j$: with CRRA utility, $v(w_j) - \delta\phi_j = \text{const}$ gives the power-law wage formula (23). Since $w_j < w_n$ at altruist-only firms, normal workers strictly prefer the marginal firm or higher-externality firms. The marginal firm index m and the normal-firm wage w_n are pinned down by market clearing ($\sum_j L_j = L$) and the altruist supply constraint ($\sum_{j=1}^{m-1} L_j \leq ML \leq \sum_{j=1}^{m} L_j$). □

Remark 7. *Altruistic sorting is the many-firm analogue of the interior equilibrium in the two-firm model (Proposition 1). Workers accept lower wages at cleaner firms because the behavioral benefit of reduced externality outweighs the consumption cost. With many firms, the sorting generates a wage ladder: the cleanest firm pays the least, each successive firm pays more, and the highest-externality firms pay the most.*

Figure 5 illustrates the altruistic sorting equilibrium with 10 firms and 200 workers.

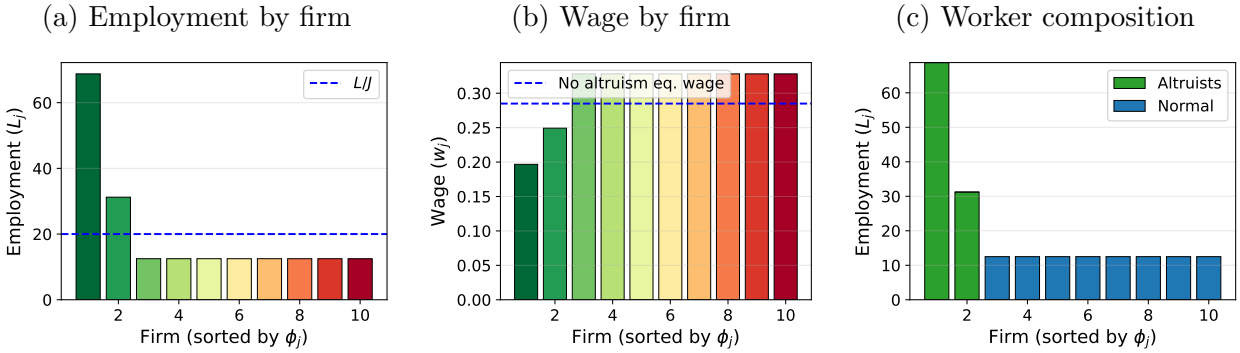
Figure 6 shows how aggregate outcomes respond to increasing altruism intensity δ in the sorting equilibrium. As δ rises, altruists pile more aggressively into clean firms, reducing the net externality (panel a) and increasing social welfare (panel b).

4.4 Deontological Altruists without a Non-Profit

The preceding analysis assumes consequentialist altruists who evaluate actions by their consequences for the net externality E. An alternative framework is *deontological* altruism: workers who follow a "do no harm" rule, refusing to work at any firm that generates externalities. Formally, deontological altruists are constrained to work only at firms with $\phi_j = 0$.[8] Like consequentialists, they maximize $v(c) - \delta E$, but subject to this employment constraint.

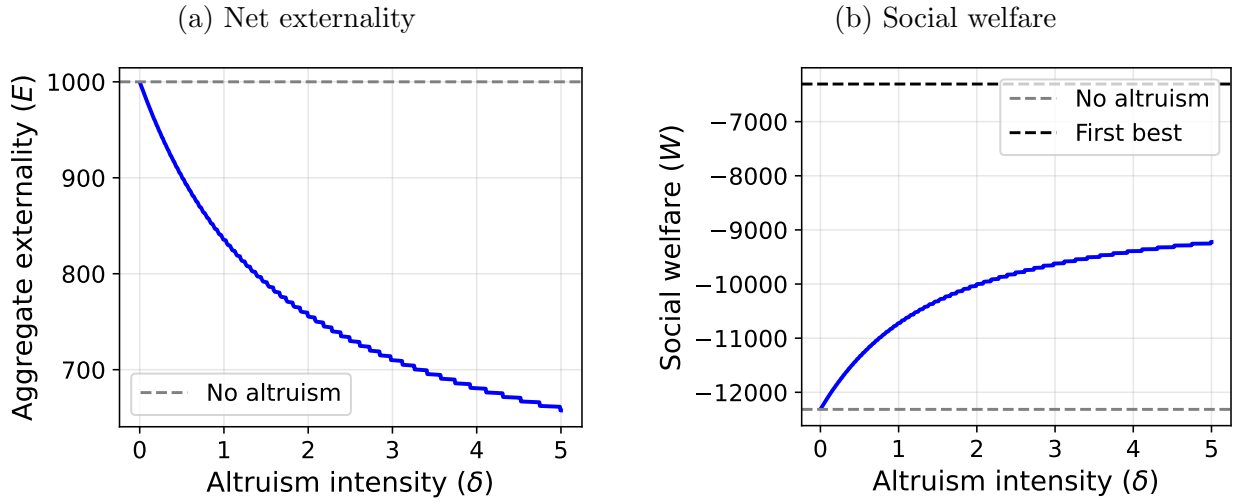
[8]The deontological constraint depends on the normalization of externality rates. I normalize $\phi_{(1)} = 0$ (the cleanest firm has zero externality), so firm 1 is the only for-profit employer that deontological altruists find acceptable. If instead the externality rates were normalized so that the worst firm had $\phi_{(J)} = 0$

Figure 5: Altruistic sorting



Notes: $J = 10$ firms with common productivity $A = 1.0$ and linearly spaced externality rates $\phi_j \in [0, 2]$. There are $L = 200$ workers with altruist share $M = 0.50$ ($\delta = 5.0$). Firms sorted by ϕ_j (ascending); bar color indicates ϕ_j (green = low, red = high). Panel (a): altruists crowd into clean firms, making them overcrowded relative to the equal-employment benchmark. Panel (b): dashed line shows the no altruism wage w^*. Clean firms pay less (overcrowded); high-externality firms pay more (undercrowded). Panel (c): green = altruists, blue = normal workers. Parameters: $\alpha = 0.7$, $\sigma = 2$.

Figure 6: Sorting equilibrium outcomes



Notes: $J = 200$ firms with common productivity $A = 1.0$ and linearly spaced externality rates $\phi_j \in [0, 2]$. Dashed gray line: no altruism equilibrium. The net externality $E = \sum_j \phi_j L_j$ falls as altruists sort more aggressively into clean firms. Social welfare $W = \sum_k v(w_k) - L\gamma E$ increases because the externality reduction outweighs the consumption cost of wage dispersion. δ ranges from 0 to $\Gamma/2 = 5.0$ ($\lambda \leq 0.5$). Parameters: $\alpha = 0.7$, $L = 1000$, $M = 0.50$, $\gamma = 0.01$, $\sigma = 2$.

Without a non-profit, the deontological equilibrium is straightforward. All ML deontological altruists work at firm 1 ($\phi_{(1)} = 0$). The $(1-M)L$ normal workers distribute equally across firms $2, \ldots, J$, equalizing wages at those firms.

Proposition 7 (Deontological Sorting without Non-Profit)**.** *With J firms, $\phi_{(1)} = 0$, and a share M of deontological altruists:*

1. *Firm 1 employs all ML altruists (and possibly some normal workers if $ML < L/J$). Firms $2, \ldots, J$ employ only normal workers.*

2. *When $M \leq 1/J$ (futility regime), normal workers are present at all firms, wages equalize, and the equilibrium is identical to the no altruism equilibrium—just as with consequentialists.*

3. *When $M > 1/J$, firm 1 is overcrowded with wage $w_1 = \alpha A(ML)^{\alpha-1} < w_n$, where $w_n = \alpha A((1-M)L/(J-1))^{\alpha-1}$ is the wage at normal-worker firms.*

4. *The net externality is $E_d = \frac{(1-M)L}{J-1} \sum_{j=2}^{J} \phi_{(j)}$, which is independent of δ.*

Proof. Parts (1)–(2) mirror Proposition 5: when $ML \leq L/J$, all altruists fit at firm 1 alongside normal workers, and wages equalize. For part (3), when $ML > L/J$, no normal workers go to firm 1 (they earn more at firms $2, \ldots, J$). Market clearing gives $L_j = (1-M)L/(J-1)$ for $j = 2, \ldots, J$. Part (4): $E_d = \sum_{j=2}^{J} \phi_{(j)} \cdot (1-M)L/(J-1)$. □

Remark 8 (Consequentialists versus Deontologists in the Sorting Equilibrium)**.** *The qualitative behavior is similar: both types avoid high-externality firms, and both leave the futility result intact when $M \leq 1/J$. The quantitative difference is in the sorting pattern when $M > 1/J$. Consequentialists spread across the m cleanest firms (Proposition 6), generating a wage ladder. Deontologists all pile into firm 1, producing a single large wage gap between firm 1 and the rest. With linearly spaced $\phi_j \in [0, \bar{\phi}]$, the net externality under both types is of the same order: both redirect altruistic labor away from high-externality firms. The main welfare difference comes from wage dispersion: deontologists create more overcrowding at firm 1 (and thus a larger consumption loss) than consequentialists, who spread the wage penalty across multiple clean firms.*

and all other firms had negative ϕ_j (representing positive externalities), the deontological constraint would not bind at all—every firm would satisfy $\phi_j \leq 0$, and deontological altruists would behave identically to consequentialists. The results in this section therefore depend on the assumption that at least some firms generate strictly positive externalities relative to the cleanest firm, which is the economically natural case. As shown below, deontological altruists may also work at the non-profit.

5 Equilibrium with Non-Profit

This section develops the third main result: the non-profit weakens assortative matching and creates a three-way trade-off for altruistic workers. I introduce a non-profit sector funded by altruistic donations. The non-profit fundamentally changes the calculus of firm choice. Without it, altruists sort into cleaner firms to reduce the externality directly. With the non-profit, this sorting motive weakens: an altruist who works at a higher-externality firm and donates part of their earnings to fund abatement may achieve a greater reduction in the externality than one who sorts into the clean firm at a lower wage. Each altruist faces a trade-off between earning to give at a higher-externality firm, working at a cleaner firm for lower pay and giving less, and working at the non-profit. In interior equilibrium, the wage ladder adjusts so that altruists are exactly indifferent between all three strategies (Proposition 9). Whether altruists contribute through labor donations at the non-profit or monetary donations from for-profit wages, the equilibrium allocation is the same (Proposition 8). The donation channel also breaks the futility result: even when altruists are too few to affect outcomes through firm choice alone, their contributions to the non-profit reduce the externality.

5.1 Altruistic Workers' Contribution Problem

Consider an altruist working at any for-profit firm j in the many-firm model. They earn w_j and can donate d to the non-profit. The key observation is that the non-profit's effective cost per worker is not the market wage w_n but the non-profit wage $w_{\text{np}} = c_a$: because altruistic workers accept below-market wages at the non-profit, each dollar donated hires $1/c_a$ non-profit workers, generating ψ/c_a units of abatement. Only altruists accept the non-profit wage $c_a < w_n$; their optimal consumption *is* c_a, so they consume their full non-profit wage and make no further monetary donations. For-profit altruists therefore see the return to \$1 donated as $\delta\psi/c_a$. Their behavioral utility is $v(w_j - d) - \delta E$. Taking the net externality as given, the first-order condition for donations is:

$$v'(c_a) = \frac{\delta\psi}{c_a}. \tag{24}$$

With CRRA utility $v(c) = c^{1-\sigma}/(1-\sigma)$, this gives $c_a^{-\sigma} = \delta\psi/c_a$, i.e., $c_a^{1-\sigma} = \delta\psi$, so:

$$c_a^* = (\delta\psi)^{-1/(\sigma-1)}, \qquad d_j^* = w_j - c_a^*. \tag{25}$$

With $\sigma = 2$: $c_a^* = 1/(\delta\psi)$. I write c_a for c_a^* when no ambiguity arises. Crucially, the altruist's target consumption c_a^* depends only on the altruism parameter δ, the non-profit productivity ψ, and the curvature of the utility function σ—it is independent of the equilibrium wage w_n. Positive donations require $w_j > c_a^*$, i.e., the for-profit wage must exceed the altruist's target consumption.

Now consider an altruist working at the non-profit. The non-profit pays $w_{\text{np}} = c_a^*$: each non-profit worker consumes their full wage and makes no further monetary donation. At $w_{\text{np}} = c_a^*$, the altruist's marginal utility of consumption exactly equals the marginal benefit of donating, so $d = 0$. Their labor directly generates ψ units of abatement.

Proposition 8 (Equivalence of Labor and Monetary Donations)**.** *Each altruist at for-profit firm j donates $d_j^* = w_j - c_a^*$ and consumes $c_a^* = (\delta\psi)^{-1/(\sigma-1)}$ regardless of employer. Each altruist's monetary donation $d^* = w_j - c_a^*$ funds d^*/c_a^* non-profit workers at wage c_a^*. An altruist working at the non-profit at $w_{np} = c_a^*$ earns c_a^*, donates nothing, and consumes c_a^*. Both channels yield the same effective contribution per unit of consumption sacrificed, and total non-profit employment is $L_{np} = ML \cdot (w^* - c_a^*)/c_a^*$ in the equal-wage case.*

Proof. An altruist at for-profit firm j earns w_j, donates $d_j^* = w_j - c_a^*$, and consumes $c_a^* = (\delta\psi)^{-1/(\sigma-1)}$. Each dollar donated hires $1/c_a^*$ non-profit workers, generating ψ/c_a^* units of abatement. An altruist at the non-profit at $w_{\text{np}} = c_a^*$ earns c_a^*, donates nothing, and consumes c_a^*; their labor directly generates ψ units of abatement. In the equal-wage case ($w_j = w^*$), each altruist's total effective contribution is $w^* - c_a^*$ regardless of whether they work at a for-profit firm or the non-profit. Total non-profit employment is $ML(w^* - c_a^*)/c_a^*$. □

The non-profit pays $w_{\text{np}} = c_a^*$, the lowest wage that altruists accept. At this wage, normal workers strictly prefer the for-profit sector ($w_n > c_a^*$ when the non-profit is active), so the non-profit naturally screens for altruistic workers. I adopt $w_{\text{np}} = c_a^*$ as the default; the equivalence result ensures this is without loss of generality in the frictionless model.[9]

Remark 9 (Linearity and the Absence of Strategic Substitutability)**.** *The externality enters altruistic utility* linearly*: $u_a = v(c) - \delta E$. This implies that the marginal benefit of donating, $\delta\psi/c_a$, is independent of what other altruists donate. In contrast, the canonical model of private provision of public goods (Bergstrom et al., 1986) features strategic substitutability: when others contribute more, each agent's marginal benefit of contributing falls (due to concavity of preferences over the public good level). Here, linearity in E eliminates this effect entirely—there is no crowding out among donors. If instead altruistic utility were*

[9]With frictions such as income taxes that are not fully offset by donation deductions, labor donations are more efficient than monetary donations: the altruist avoids paying tax on d^* and then donating the after-tax remainder. The below-market wage formulation is therefore more robust to institutional details.

$v(c) - \delta \cdot h(E)$ with h strictly convex, donations would be strategic substitutes and the Warr (1983) neutrality result would apply.[10]

5.2 The Non-Profit Weakens Altruistic Sorting

Section 4.3 shows that without the non-profit, altruists sort into clean firms, accepting lower wages in exchange for a reduced externality. The sorting condition is $w_j = [w_n^{1-\sigma} + (1-\sigma)\delta(\phi_{(m)} - \phi_{(j)})]^{1/(1-\sigma)}$: wages at altruist firms decrease with the externality differential according to a power-law relationship, scaled by δ. I now show that the non-profit fundamentally changes this sorting pattern.

With the non-profit, each altruist at firm j earns w_j, donates $d_j = w_j - c_a$ where $c_a = (\delta\psi)^{-1/(\sigma-1)}$, and consumes c_a regardless of their employer. Their behavioral utility from working at firm j and donating optimally includes (i) the consumption component $v(c_a)$ (the same at all firms), (ii) the externality cost $-\delta\phi_j$ from their own labor at firm j, and (iii) the abatement benefit $\delta\psi(w_j - c_a)/c_a$ from their donation (each dollar donated hires $1/c_a$ NP workers). Equalizing across firms:

$$-\delta\phi_j + \frac{\delta\psi w_j}{c_a} = \text{const}, \tag{26}$$

which gives the *modified sorting condition*:

$$w_j = w_n - \frac{c_a(\phi_{(m)} - \phi_{(j)})}{\psi}, \qquad j = 1, \ldots, m-1. \tag{27}$$

Proposition 9 (Non-Profit Weakens Altruistic Sorting)**.** *With J firms of common productivity A, a share M of altruists with $w_n > c_a^*$, and a non-profit funded by donations:*

1. *The sorting condition changes from the CRRA power-law* (23) *to linear* (27)*: the wage gap at altruist firms is $c_a(\phi_{(m)} - \phi_{(j)})/\psi$, with slope c_a/ψ.*

2. *Sorting now depends on δ through $c_a = (\delta\psi)^{-1/(\sigma-1)}$, but weakly: higher δ lowers c_a, producing a flatter wage ladder. The donation channel partially absorbs the sorting motive—more altruistic workers donate more and sort less aggressively.*

3. *Since $c_a < w_n$ whenever the non-profit is active, the sorting coefficient c_a/ψ is smaller than the no-NP sorting strength: the non-profit always weakens concentration relative*

[10]The Warr (1983) neutrality result states that the equilibrium level of a voluntarily provided public good is independent of the income distribution among contributors. In the present model, total non-profit employment depends on total effective contributions, not on whether contributions take the form of monetary donations or labor at below-market wages (Proposition 8).

to the case without it.

4. *As $\psi \to \infty$, $c_a \to 0$ and the wage gap vanishes: sorting dissolves completely and the for-profit allocation converges to the no altruism equilibrium plus a non-profit sector funded by donations.*

Proof. Part (1) follows from the modified FOC. Without the non-profit, the altruist's firm-choice condition is $v(w_j) - \delta\phi_j = \text{const}$, giving the power-law formula (23) with CRRA utility. With the non-profit, the altruist's net benefit from firm j (after optimal donations) includes the abatement term $\delta\psi w_j/c_a$: the indifference condition becomes $-\delta\phi_j + \delta\psi w_j/c_a = \text{const}$, yielding the linear condition (27). Part (2): δ enters through $c_a = (\delta\psi)^{-1/(\sigma-1)}$; higher δ lowers c_a and thus the slope c_a/ψ. Part (3): the non-profit is active when $w_n > c_a$, so the sorting coefficient $c_a/\psi < w_n/\psi$, which is weaker than the no-NP sorting. Part (4): as $\psi \to \infty$, $c_a = (\delta\psi)^{-1/(\sigma-1)} \to 0$, so $c_a/\psi \to 0$ and all wages converge to w_n. □

Remark 10 (Consequentialist versus Deontological Altruism)**.** *While Section 4.4 shows that consequentialist and deontological altruists produce similar outcomes in the sorting equilibrium, the non-profit introduces a sharp divergence. Consequentialist altruists may rationally work at a high-externality firm and donate the wage premium to the non-profit—an earn-to-give strategy that can reduce the externality by more than sorting into a clean firm. Deontological altruists who refuse employment at any firm with $\phi_j > 0$ cannot exploit this channel, leading to lower donations and higher aggregate externalities (Section 5.6).*

Figure 7 illustrates how the non-profit weakens concentration.[11] With the non-profit (red), the wage differential between clean and high-externality firms is smaller than without (blue): altruists spread more evenly across firms, channeling the wage premium at higher-externality firms into donations rather than into additional overcrowding at clean firms. The green bar labeled "NP" in panels (a) and (b) shows the non-profit wage c_a^* and employment L_{np}, respectively. Total for-profit employment is lower with the non-profit ($L_{\text{priv}} = L - L_{\text{np}} < L$): the workers who move to the non-profit reduce total for-profit production and thus total harm $H = \sum_j \phi_j L_j$, even though per-firm sorting weakens.

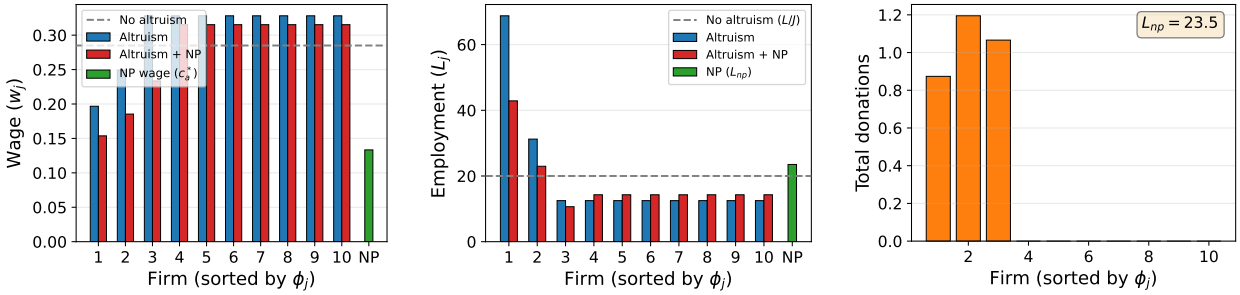
5.3 Equilibrium with Non-Profit

When $ML \leq L_{\text{priv}}/J$—or when ψ is large enough that the non-profit dissolves sorting (Proposition 9, part 4)—all for-profit firms have equal employment and a common wage w^*. The

[11] The jump in employment at the boundary between altruist and normal firms (visible in panel b between firms 3 and 4) reflects the sharp labor-market segmentation that arises under perfect competition. Firms paying below w_n attract only altruists; firms paying w_n attract only normal workers. The marginal firm m can employ both types. With search frictions or imperfect information, the transition would be smoother.

Figure 7: Non-profit weakens altruistic sorting

(a) Wage by firm (b) Employment by firm (c) Total donations by firm



Notes: Same parameters as Figure 5 ($J = 10$, $L = 200$, $M = 0.50$, $\delta = 5.0$, $\alpha = 0.7$, $\phi_j \in [0, 2]$), with non-profit productivity $\psi = 1.5$. Blue bars: altruism (no non-profit). Red bars: altruism with non-profit. Green bar ("NP"): non-profit wage c_a^* (panel a) and non-profit employment L_{np} (panel b). Gray dashed line: no altruism equilibrium. The non-profit weakens sorting: altruists spread out more, earning higher wages at higher-externality firms and donating the premium to the non-profit. Total for-profit employment is lower with the non-profit because $L_{\text{priv}} = L - L_{\text{np}} < L$. $\sigma = 2$.

closed-form expressions below apply in this equal-wage case; the figures in this section compute the full sorting equilibrium numerically and are not restricted to it. In the equal-wage case, each altruist donates $d^* = w^* - c_a^*$ where $c_a^* = (\delta\psi)^{-1/(\sigma-1)}$, so total non-profit employment is

$$L_{\text{np}} = ML \cdot \frac{w^* - c_a^*}{c_a^*}, \tag{28}$$

where $w^* = \alpha A(L_{\text{priv}}/J)^{\alpha-1}$ depends on L_{np} through $L_{\text{priv}} = L - L_{\text{np}}$. Note that L_{np} cannot be written in closed form independent of w^*, because the donation $d^* = w^* - c_a^*$ depends on the equilibrium wage. However, since c_a^* is determined by parameters alone, L_{np} is the solution to a single equation in one unknown. The equilibrium is characterized by:

1. Non-profit employment: L_{np} solves $L_{\text{np}} = ML(w^* - c_a^*)/c_a^*$ jointly with $w^* = \alpha A((L - L_{\text{np}})/J)^{\alpha-1}$, if $w^* > c_a^*$; else $L_{\text{np}} = 0$.

2. For-profit employment: $L_{\text{priv}} = L - L_{\text{np}}$.

3. Allocation across for-profit firms: $L_j = L_{\text{priv}}/J$ for each j.

4. Wage: $w^* = \alpha A(L_{\text{priv}}/J)^{\alpha-1}$.

The net externality is:

$$E = \frac{L_{\text{priv}}}{J} \sum_j \phi_j - \psi L_{\text{np}} = L_{\text{priv}}\bar{\phi} - \psi L_{\text{np}}, \tag{29}$$

where $\bar{\phi} = (1/J)\sum_j \phi_j$ is the average externality rate.

Proposition 10 (Altruism Reduces the Externality)**.** *In the equal-wage case ($ML \leq L_{priv}/J$), for $w^* > c_a^*$:*

1. *The net externality E is strictly decreasing in δ (altruism intensity) and in M (altruist share).*

2. *Non-profit employment L_{np} and total effective contributions are strictly increasing in δ and M.*

3. *The equilibrium wage w^* is strictly increasing in δ and M (fewer workers competing for for-profit jobs raises the marginal product).*

Proof. Higher δ lowers $c_a^* = (\delta\psi)^{-1/(\sigma-1)}$, increasing $d^* = w^* - c_a^*$ and thus L_{np}. From (29), $\partial E/\partial L_{\text{np}} = -[\bar{\phi} + \psi] < 0$, so $\partial E/\partial\delta < 0$. Parts for M are analogous. For part (3), $w^* = \alpha A[(L - L_{\text{np}})/J]^{\alpha-1}$, and since $\alpha - 1 < 0$, decreasing $L - L_{\text{np}}$ increases w^*. □

This result contrasts sharply with Proposition 2: with a non-profit, altruism has bite. The channel is direct—altruistic contributions fund abatement workers—and is not undone by general equilibrium substitution in the for-profit labor market.

5.4 Who Works at the Non-Profit?

Remark 11 (Earn to Give)**.** *At $w_{np} = c_a^*$, only altruists accept employment at the non-profit. In the equal-wage case, Proposition 8 shows that the equilibrium allocation is the same if the non-profit instead pays w^* and relies on monetary donations: the identity of non-profit workers is irrelevant, just as the identity of workers at the for-profit firms is irrelevant. What matters for the externality is who* contributes*, not who* works *at the non-profit. This formalizes the earn-to-give logic: an altruist generates exactly the same abatement by earning w^* at a for-profit firm and donating d^* as by working directly at the non-profit at $w_{np} = c_a^*$. Morduch and Szafarz (2019) study the same trade-off in partial equilibrium and show that equivalence breaks down when workers face binding budget constraints: constrained altruists are better off donating labor, while unconstrained altruists should earn to give. In the present model, a different source of non-equivalence arises: when wages differ across firms (the sorting case), monetary and labor donations generate different levels of non-profit funding (Proposition 8). In either case, in the presence of frictions that make monetary donations costly—such as income taxes without full donation deductibility—the below-market wage channel dominates: the non-profit should hire altruists at $w_{np} = c_a^*$ rather than paying market wages and relying on donations.*

5.5 Welfare with Non-Profit

Proposition 11 (Welfare with Non-Profit)**.** *In the equal-wage case, social welfare is* $W = (1 - M)L\,v(w^*) + ML\,v(c_a^*) - L\,\gamma\,E$*, where* $v(c) = c^{1-\sigma}/(1-\sigma)$ *and* $c_a^* = (\delta\psi)^{-1/(\sigma-1)}$*. There exist parameter values such that introducing the non-profit (moving from* $d^* = 0$ *to the interior optimum) strictly increases social welfare.*

Proof. When w^* is just above $c_a^* = (\delta\psi)^{-1/(\sigma-1)}$, d^* is close to zero, so the consumption loss is second-order. But the reduction in E is first-order (from the marginal non-profit worker's abatement). For γ sufficiently large, the externality reduction dominates. □

Remark 12 (Welfare Monotonicity under Bounded Altruism (Approximate Model))**.** *In the approximate model where behavioral utility is* $v(c) - \delta E$*, when* $\delta \leq L\gamma$ *(equivalently* $\lambda \leq 1$*; see Section 2.4), social welfare is monotonically increasing in* δ*. The intuition is as follows. At* $\delta = L\gamma$*, the altruist's marginal valuation of the externality equals the true social marginal harm, so the marginal incentive to donate is aligned with the social cost:* $v'(c_a^*) = L\gamma\psi/c_a^*$*. For any* $\delta < L\gamma$*, altruists undervalue the externality and therefore under-donate relative to this benchmark. Increasing* δ *moves donations toward the benchmark, and welfare rises. Only if* $\delta > L\gamma$ *(altruists "over-caring" relative to the true social harm) could welfare decrease. In the exact utilitarian model* (10)*, where the altruist's objective includes the consumption externality terms* $\lambda \sum_{k \neq j} v(c_k)$*, welfare is monotonically increasing in* λ *for all* $\lambda \in [0, 1]$*; any welfare decline at extreme values of* δ *in the approximate model is an artifact of the approximation that drops these terms.*[12]

Remark 13 (Altruists Are Worse Off with Non-Profit)**.** *With a non-profit, altruistic workers' experienced utility* $v(c_a^*) - \gamma E$ *differs from normal workers' experienced utility* $v(w^*) - \gamma E$ *because* $c_a^* = (\delta\psi)^{-1/(\sigma-1)} < w^*$*. Altruists are worse off in experienced terms, but this is because they voluntarily choose to donate—reducing their own consumption to fund abatement—not because of heightened awareness. This stands in contrast to the no altruism case (Proposition 2), where altruists' experienced welfare is identical to normal workers'.*

Figure 8 illustrates the non-profit equilibrium in the many-firm model, comparing outcomes with and without the non-profit as altruism intensity varies. Panels (a) and (b) decompose the net externality into its two components: the total harm $H = \sum_j \phi_j L_j$ from for-profit production and the total good $G = \psi L_{\text{np}}$ from non-profit abatement. Without a non-profit (solid blue), only sorting reduces H by redirecting labor toward cleaner firms.

[12]This argument takes the equilibrium wage w^* as given. A social planner would additionally internalize the effect of non-profit employment on for-profit wages, so the planner's optimum does not exactly coincide with the altruist's donation FOC even at $\delta = L\gamma$. The monotonicity conclusion holds because the direction of the effect is unambiguous: at $\delta < L\gamma$, altruists undervalue the externality.

With a non-profit (dashed red), total harm H is *lower* despite weaker sorting because the non-profit drains workers from the for-profit sector ($L_{\text{priv}} = L - L_{\text{np}} < L$), reducing total for-profit employment and thus total harm. The non-profit additionally generates $G > 0$ (panel b), so the combined effect on the net externality $E = H - G$ (panel c) is substantially larger than sorting alone. Panel (d) shows that the non-profit amplifies the welfare gains from altruism. Figure A.1 in Appendix A shows the corresponding results for the two-firm special case where the futility result holds without a non-profit.

The effectiveness of the non-profit (ψ) governs the trade-off between two equilibrium channels. When $\psi = 0$, the non-profit is inactive and only sorting matters: altruists concentrate at clean firms, accepting lower wages to reduce the externality directly. As ψ increases, the donation channel becomes viable and eventually dominates. In the limit, altruists earn the same wages as normal workers—the sorting motive dissolves—and externality reduction comes entirely through non-profit abatement funded by donations. Figure 9 traces this transition. When ψ is low (orange line), the non-profit barely improves upon pure sorting. For moderate ψ (green), donations meaningfully supplement sorting. When ψ is high (red), the non-profit dominates and drives large reductions in E. Crucially, the non-profit only becomes active once the for-profit wage exceeds the altruist's target consumption $w_n > c_a^* = (\delta\psi)^{-1/(\sigma-1)}$, below which the marginal return to donating is insufficient to justify any sacrifice in consumption.

5.6 Deontological Altruists with a Non-Profit

Section 4.4 shows that without a non-profit, deontological and consequentialist altruists produce qualitatively similar outcomes: both avoid high-externality firms, and the main difference is in the distribution of altruists across clean firms. With a non-profit, however, the two frameworks diverge sharply.

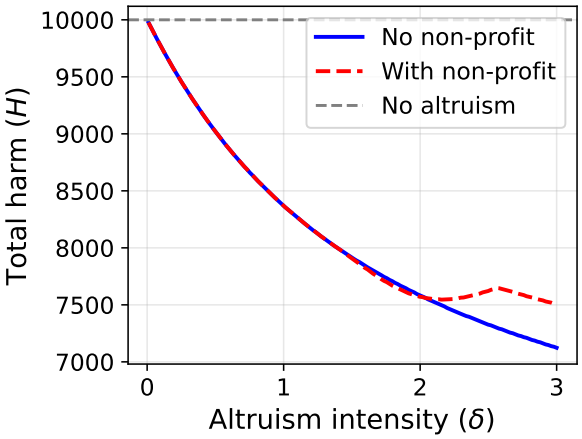
Recall that consequentialist altruists may work at any firm and donate the wage premium to the non-profit (Proposition 9). The earn-to-give strategy exploits the fact that wages are higher at high-externality firms (which are undercrowded), so consequentialists working there generate larger donations and fund more abatement. Deontological altruists, by contrast, refuse to work at any firm with $\phi_j > 0$. They can only work at firm 1 ($\phi_{(1)} = 0$) or at the non-profit, and can only donate from the wage earned at firm 1.

Proposition 12 (Deontological Altruists with Non-Profit)**.** *With J firms, $\phi_{(1)} = 0$, a share M of deontological altruists, and a non-profit with productivity ψ:*

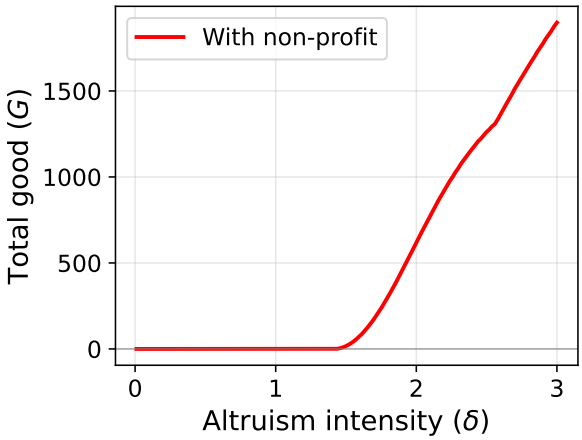
1. *All altruists work at firm 1 or at the non-profit. Their wage at firm 1 is $w_1 =$*

Figure 8: Non-profit sector (many firms)

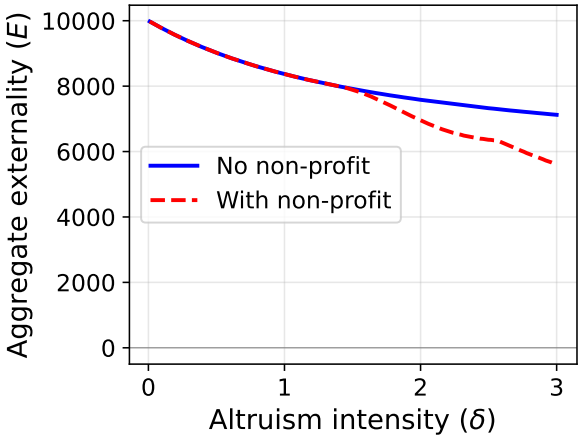
(a) Total harm (H)



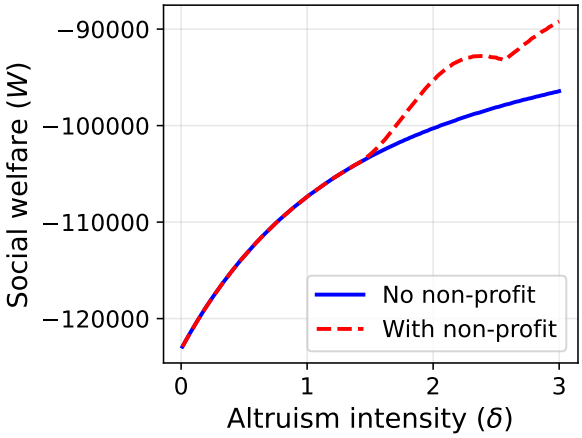
(b) Total good (G)



(c) Net externality ($E = H - G$)



(d) Social welfare

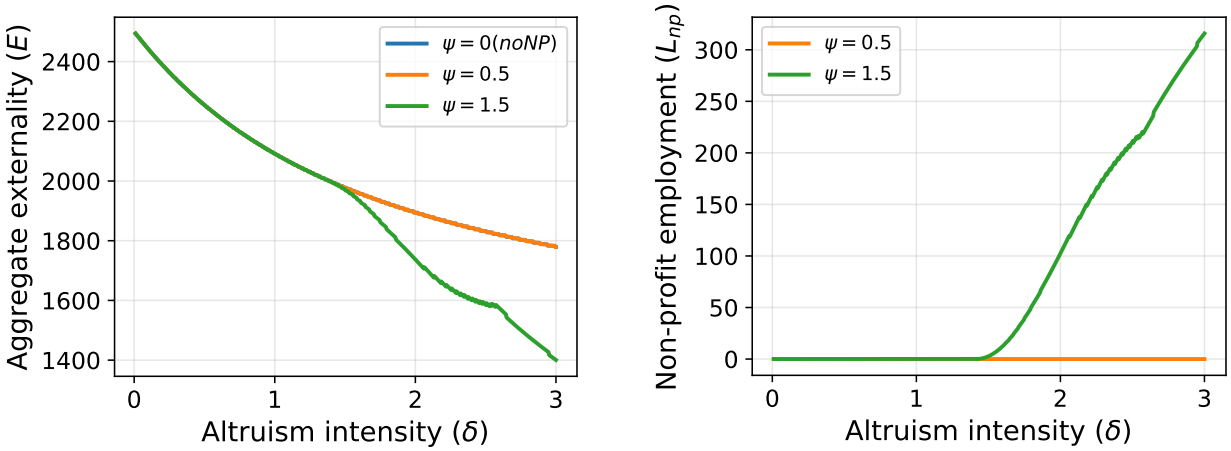


Notes: Equilibrium outcomes as altruism δ varies, in the many-firm model ($J = 1000$). Solid blue: without non-profit (sorting equilibrium only). Dashed red: with non-profit ($\psi = 1.5$). Panels (a) and (b) decompose the net externality into total harm $H = \sum_j \phi_j L_j$ and total good $G = \psi L_{\text{np}}$. Panel (c) shows the net externality $E = H - G$. Panel (d) shows social welfare. δ ranges from 0 to 0.3Γ ($\lambda \leq 0.3$); Appendix D discusses why the behavioral utility approximation becomes less accurate at higher δ. Parameters: $A = 1.0$, $\alpha = 0.7$, $\phi_j \in [0, 2]$, $L = 5{,}000$, $M = 0.50$, $\gamma = 0.01$, $\sigma = 2$.

Figure 9: Non-profit effectiveness and the sorting–donation trade-off

(a) Net externality

(b) Non-profit employment



Notes: Net externality E and non-profit employment L_{np} as altruism intensity δ varies, for different values of non-profit effectiveness ψ. At $\psi = 0$ (no non-profit), only sorting reduces E. As ψ increases, the donation channel strengthens. The non-profit becomes active when $w_n > c_a^* = (\delta\psi)^{-1/(\sigma-1)}$; higher ψ lowers c_a^* and activates the non-profit at lower δ. δ ranges from 0 to $0.3\Gamma = 3.0$ ($\lambda \leq 0.3$). Parameters: $J = 500$, $A = 1.0$, $\alpha = 0.7$, $\phi_j \in [0, 2]$, $L = 2500$, $M = 0.50$, $\gamma = 0.004$, $\sigma = 2$.

$\alpha A(M_1)^{\alpha-1}$, *where* M_1 *is the number of altruists at firm 1. Normal workers equalize wages across firms* $2, \ldots, J$ *at* $w_n = \alpha A((1-M)L/(J-1))^{\alpha-1}$.

2. *Each altruist at firm 1 donates* $d_1 = w_1 - c_a^*$ *(when* $w_1 > c_a^*$*), consuming* $c_a^* = (\delta\psi)^{-1/(\sigma-1)}$ *—the same consumption as consequentialist altruists.*

3. *When* $M > 1/J$*, total donations per altruist are lower than under consequentialism:* $d_1 = w_1 - c_a^* < w_n - c_a^*$*, because* $w_1 < w_n$ *(firm 1 is overcrowded). A consequentialist altruist working at a normal-wage firm would donate* $d_n = w_n - c_a^* > d_1$*. (When* $M \leq 1/J$*,* $w_1 = w_n$ *and the two coincide.)*

4. *When* $M > 1/J$*, the net externality is higher than under consequentialism: the non-profit receives less funding, generating less abatement.*

Proof. Parts (1) and (2) follow from the deontological constraint ($\phi_j = 0$ only) and the standard donation first-order condition (24). For part (3), the wage at firm 1 satisfies $w_1 < w_n$ when $M > 1/J$ (overcrowding), so $d_1 = w_1 - c_a < w_n - c_a = d_n$. Total deontological effective contributions are $ML \cdot d_1 = ML(w_1 - c_a)$. Under consequentialism with the non-profit, the modified sorting condition (27) spreads altruists across multiple firms, avoiding the severe overcrowding at firm 1 that depresses w_1. Since consequentialist altruists access

higher average wages, their total effective contributions are strictly larger. Part (4): lower total contributions under deontology imply lower non-profit employment and less abatement, so $E_d > E_c$. □

The welfare cost of deontological altruism becomes transparent: by refusing to work at high-externality firms, deontological altruists forgo the wage premium that funds abatement through the non-profit. In the limit of very productive abatement ($\psi \to \infty$), the consequentialist equilibrium has all firms equally staffed and externality reduction coming entirely through donations. The deontological equilibrium, by contrast, retains overcrowding at firm 1 and underfunds the non-profit—the "do no harm" constraint prevents altruists from exploiting the earn-to-give channel. The irony is that by avoiding personal involvement in harmful production, deontological altruists allow more aggregate harm to persist.

6 Discussion and Conclusion

This paper develops a general equilibrium framework for studying altruistic labor supply, establishing three main results.

First, a benchmark futility result: even though altruists optimally sort into the cleaner firm, when they are few this sorting has no effect on outcomes. General equilibrium substitution by normal workers undoes the reallocation, and the intensity of altruism does not affect social welfare.

Second, this futility is fragile. When firms are homogeneous except in their externality rates, the threshold for futility is the employment share of a single firm, which shrinks as $1/J$—even when total employment grows proportionally. This generates a stark asymmetry between concentration and boycotting: a small group of altruists can break through futility by concentrating at the best employer long before they could affect outcomes by boycotting the worst. The practical implication is that for a small group of motivated workers, identifying the best employer matters far more than identifying the worst. This supports a concentration strategy—as advocated by organizations like 80,000 Hours (Todd, 2023), which recommend high-impact career paths—over a boycotting strategy that urges workers to avoid the worst employers. When altruists are numerous enough to exceed the cleanest firm's capacity, the equilibrium features assortative matching: altruists concentrate at the cleanest firms, accepting lower wages, while high-externality firms become undercrowded and pay higher wages, generating a compensating differential schedule across firms. Social welfare is increasing in the intensity of altruism.

Third, the non-profit weakens assortative matching. The sorting condition shifts from a

power-law in the altruism parameter δ to linear with slope c_a/ψ that depends only weakly on δ, as the donation channel partially absorbs the sorting motive. Individual altruists face a three-way trade-off—between earning to give at a higher-externality firm, working at a cleaner firm for lower pay, and working at the non-profit—and in interior equilibrium are indifferent between all three strategies. The non-profit always weakens sorting relative to the case without it, and in the limit of very productive abatement ($\psi \to \infty$), sorting dissolves completely. This formalizes the "earn to give" logic of the effective altruism movement: altruistic workers generate social value through their willingness to donate, not through their choice of employer.

The comparison between consequentialist and deontological altruism sharpens the earn-to-give insight. Without a non-profit, the two ethical frameworks produce similar outcomes: both lead altruists to avoid high-externality firms. With a non-profit, the frameworks diverge. Consequentialists exploit the earn-to-give channel—working at high-externality firms for higher wages and donating the premium—while deontologists refuse employment at harmful firms and therefore donate less. The result is lower social welfare under deontological altruism, even though deontologists are not personally engaged in harmful activity.

The model's earn-to-give result resonates with a strategy that was prominent in the early effective altruism movement (MacAskill, 2015): choose a high-paying career and donate a large share of income to effective charities. More recently, this strategy is no longer promoted as one of the best ways to do as much good as possible. 80,000 Hours, for example, now advises against pursuing a career in which the direct effects of the work are seriously harmful, even if the overall benefits seem greater—and argues that this position is justified even from a purely consequentialist perspective (80,000 Hours, 2025). The reasons include moral uncertainty, hidden costs to reputation and personal character, systematic overestimation of the benefits of one's own actions, and the availability of similarly paid roles without the associated harm (Todd and MacAskill, 2017). All careers involve some degree of negative impact, but 80,000 Hours recommends avoiding actions that seem clearly wrong from a commonsense perspective, even when the expected net benefit appears positive. This position has evolved significantly: early EA arguments emphasized that the harm from working at, say, a financial firm was likely small because a replacement worker would take the position anyway (MacAskill, 2014), but the organization later walked back the strength of this replaceability argument (80,000 Hours, 2015). Singer (2015) and MacAskill (2015) both discussed earning to give on Wall Street favorably, but by 2017 the consensus had shifted toward caution. The collapse of FTX further reinforced the case for commonsense ethical constraints on career choice (80,000 Hours, 2023).

A recurring theme in philosophical ethics is whether one may do something harmful in

order to produce a greater good—as in Thomson's transplant problem, where a surgeon could kill one healthy patient to harvest organs that save five others (Thomson, 1985).[13] Critics of effective altruism have invoked this structure in the wake of the FTX scandal, arguing that the earn-to-give logic inevitably leads people to engage in activities that most would agree are wrong—financial fraud, environmental destruction, exploitative labor practices—in order to fund donations (80,000 Hours, 2023). The concern is that consequentialist reasoning, taken to its logical extreme, licenses any means so long as the ends are sufficiently beneficial.

The model provides its own, purely economic rationale for why this extreme scenario rarely arises in practice. When all workers have the same productivity, even if a single firm has an outlier-large externality, normal workers are drawn to that firm by its high wages, bidding the wage down until it offers no significant premium over less harmful alternatives. In equilibrium, high-externality firms pay the same wage w_n as other normal-worker firms. Only when altruists are numerous enough to trigger sorting (Proposition 6) do compensating differentials emerge—and even then, it is the clean firms that deviate from w_n (paying less), not the dirty ones (Appendix C). The altruist's earn-to-give choice is therefore between similar-wage firms, not between an ethically clean low-paying job and a highly harmful high-paying one.

Moreover, when workers have heterogeneous altruism and all workers are at least slightly altruistic, the model generates an even stronger result: the most harmful activities *do* pay a wage premium (because few workers are willing to do them), but it is precisely the *least* altruistic workers who fill those positions, not the most altruistic ones. A "hardcore" altruist with high δ would never find it optimal to work at the outlier-harmful firm, because the sorting condition ensures that the behavioral cost of the externality ($\delta\phi_J$) far exceeds any wage premium. The workers who might accept employment at such a firm are those with the lowest δ—workers who care least about externalities. Financial fraud, for instance, may pay well precisely because few are willing to engage in it, but the model predicts that a highly altruistic person would lose more in behavioral utility from the extreme externality than they would gain from the wage premium and subsequent donations. Appendix C formalizes this argument. The model thus reinforces the practical advice of 80,000 Hours and the broader effective altruism community: the scenario in which a large personal sacrifice of ethics is required to fund large donations is more of a philosophical thought experiment than a realistic trade-off (Todd, 2023; 80,000 Hours, 2025).

[13]Thomson (1985) introduced the transplant problem to contrast it with the trolley problem. In the transplant case, a surgeon has five patients each dying from a different organ failure. A healthy young man comes in for a routine checkup, and his organs are compatible with all five. May the surgeon kill the one to save the five? Most people say no, even though the utilitarian calculus favors the transplant—illustrating the limits of pure consequentialism.

The present model abstracts from moral uncertainty, reputational effects, and scope for self-serving bias, and therefore isolates the pure consequentialist logic of earn-to-give. The gap between this theoretical benchmark and the practical advice of the effective altruism movement underscores the importance of the considerations that the model omits.

Several extensions are left for future work. Allowing firms to differ in productivity (rather than only in externality rates) would generate variation in employment shares across firms and richer sorting patterns; Appendix B sketches the main results for this case. Introducing a continuous distribution of altruism intensity would produce assortative matching rather than the binary sorting studied here. Labor market frictions—such as search costs or occupational barriers—would break the futility result by preventing the costless reallocation of normal workers. The interaction between altruistic labor supply and altruistic consumption deserves further investigation. Appendix A sketches additional directions.

References

80,000 Hours, "'Replaceability' Isn't as Important as You Might Think (Or We've Suggested)," `https://80000hours.org/2015/07/replaceability-isnt-as-important-as-you-might-think-or-weve-suggested/` 2015. Accessed March 2026.

_ , "How 80,000 Hours Has Changed Some of Our Advice After the Collapse of FTX," `https://80000hours.org/2023/05/how-80000-hours-has-changed-some-of-our-advice-after-the-collapse-of-ftx/` 2023. Accessed March 2026.

_ , "Should You Ever Work in a Harmful Career for the Greater Good?," `https://80000hours.org/articles/harmful-career/` 2025. Accessed March 2026.

Alexandrie, Gustav, "Classical Utilitarianism Revisited," 2025. Working paper, University of Zurich.

Andreoni, James, "Privately Provided Public Goods in a Large Economy: The Limits of Altruism," *Journal of Public Economics*, 1988, *35* (1), 57–73.

_ , "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving," *Economic Journal*, 1990, *100* (401), 464–477.

Arnold, Lutz G., "On the Neutrality of Socially Responsible Investing," *Theoretical Economics*, 2023, *18* (3), 1255–1285.

Ash, Michael and James K. Boyce, "Toxic 100 Air Polluters Index and Greenhouse 100 Polluters Index," Report, Political Economy Research Institute, University of Massachusetts Amherst 2024.

Ashraf, Nava, Oriana Bandiera, Edward Davenport, and Scott S. Lee, "Losing Prosociality in the Quest for Talent? Sorting, Selection, and Productivity in the Delivery of Public Services," *American Economic Review*, 2020, *110* (5), 1355–1394.

Bénabou, Roland and Jean Tirole, "Incentives and Prosocial Behavior," *American Economic Review*, 2006, *96* (5), 1652–1678.

Bentham, Jeremy, *An Introduction to the Principles of Morals and Legislation*, London: T. Payne and Son, 1789.

Bergstrom, Theodore, Lawrence Blume, and Hal Varian, "On the Private Provision of Public Goods," *Journal of Public Economics*, 1986, *29* (1), 25–49.

Besley, Timothy and Maitreesh Ghatak, "Competition and Incentives with Motivated Agents," *American Economic Review*, 2005, *95* (3), 616–636.

Brekke, Kjell Arne and Karine Nyborg, "Attracting Responsible Employees: Green Production as Labor Market Screening," 2008. Resource and Energy Economics, 30(4), 509–526.

Burbano, Vanessa C., "Social Responsibility Messages and Worker Wage Requirements: Field Experimental Evidence from Online Labor Marketplaces," *Organization Science*, 2016, *27* (4), 1010–1028.

Cassar, Lea, "Job Mission as a Substitute for Monetary Incentives: Benefits and Limits," *Management Science*, 2019, *65* (2), 896–912.

Cornes, Richard and Todd Sandler, *The Theory of Externalities, Public Goods, and Club Goods*, 2nd ed., Cambridge University Press, 1996.

Delfgaauw, Josse and Robert Dur, "Signaling and Screening of Workers' Motivation," *Journal of Economic Behavior & Organization*, 2007, *62* (4), 605–624.

Fehrler, Sebastian and Michael Kosfeld, "Pro-Social Missions and Worker Motivation: An Experimental Study," *Journal of Economic Behavior & Organization*, 2014, *100*, 99–110.

François, Patrick, "'Public Service Motivation' as an Argument for Government Provision," *Journal of Public Economics*, 2000, *78* (3), 275–299.

Gregg, Paul, Paul A. Grout, Anita Ratcliffe, Sarah Smith, and Frank Windmeijer, "How Important is Pro-Social Behaviour in the Delivery of Public Services?," *Journal of Public Economics*, 2011, *95* (7–8), 758–766.

Handy, Femida and Eliakim Katz, "The Wage Differential between Nonprofit Institutions and Corporations: Getting More by Paying Less?," *Journal of Comparative Economics*, 1998, *26* (2), 246–261.

Hedblom, Daniel, Brent R. Hickman, and John A. List, "Toward an Understanding of Corporate Social Responsibility: Theory and Field Experimental Evidence," 2019. NBER Working Paper No. 26222.

Kaufmann, Marc, Peter André, and Botond Kőszegi, "Understanding Markets with Socially Responsible Consumers," *The Quarterly Journal of Economics*, 2024, *139* (3), 1989–2035.

Kotchen, Matthew J., "Green Markets and Private Provision of Public Goods," *Journal of Political Economy*, 2006, *114* (4), 816–845.

Krueger, Philipp, Daniel Metzger, and Jiaxin Wu, "The Sustainability Wage Gap," 2024. Swiss Finance Institute Research Paper No. 21-17.

MacAskill, William, "Replaceability, Career Choice, and Making a Difference," *Ethical Theory and Moral Practice*, 2014, *17* (2), 269–283.

_ , *Doing Good Better: How Effective Altruism Can Help You Help Others, Do Work that Matters, and Make Smarter Choices about Giving Back*, Avery, 2015.

Mill, John Stuart, *Utilitarianism*, London: Parker, Son, and Bourn, 1863.

Morduch, Jonathan and Ariane Szafarz, "Altruism and Occupational Choice," 2019. Working paper, New York University.

Preston, Anne E., "The Nonprofit Worker in a For-Profit World," *Journal of Labor Economics*, 1989, *7* (4), 438–463.

Sidgwick, Henry, *The Methods of Ethics*, London: Macmillan, 1874.

Singer, Peter, *The Most Good You Can Do: How Effective Altruism Is Changing Ideas About Living Ethically*, New Haven: Yale University Press, 2015.

Thomson, Judith Jarvis, "The Trolley Problem," *Yale Law Journal*, 1985, *94* (6), 1395–1415.

Todd, Benjamin, *80,000 Hours: Find a Fulfilling Career That Does Good*, 2nd ed., Oxford: 80,000 Hours, 2023.

_ **and William MacAskill**, "Is It Ever OK to Take a Harmful Job in Order to Do More Good? An In-Depth Analysis," `https://80000hours.org/articles/harmful-career/` 2017. 80,000 Hours. First published June 2017; last updated March 2023.

Trammell, Philip, "Altruism in General Equilibrium," 2025. Working paper, available at SSRN.

Ľuboš Pástor, Robert F. Stambaugh, and Lucian A. Taylor, "Sustainable Investing in Equilibrium," *Journal of Financial Economics*, 2021, *142* (2), 550–571.

Warr, Peter G., "The Private Provision of a Public Good is Independent of the Distribution of Income," *Economics Letters*, 1983, *13* (2–3), 207–211.

A Potential Extensions

Several directions for extending the model merit investigation:

- **Tax and redistribution.** Introduce a leisure-labor tradeoff and heterogeneous worker abilities. A social planner can tax labor income and redistribute. Key questions: does the futility result survive with distortionary taxes? With heterogeneous abilities, more productive workers may generate more externality at the bad firm, making sorting relevant. The interaction between optimal taxation and altruistic labor supply is unexplored.

- **Worker-firm match quality.** If worker productivity varies across firms (match-specific human capital), workers are no longer perfect substitutes. Altruistic workers who sort into the good firm at a wage penalty are not costlessly replaced by equally productive normal workers. This breaks the futility result and introduces a tradeoff between allocative efficiency and externality reduction.

- **Search frictions and matching.** With search costs or matching frictions, workers cannot costlessly switch firms. The futility result relies on frictionless reallocation; introducing frictions would give altruistic sorting real effects, but at the cost of reduced output from misallocation.

- **A general futility of costless altruism theorem.** The futility result in this paper arises because altruistic workers are perfect substitutes for normal workers: when an altruist vacates a position at a high-externality firm, a normal worker fills it at no cost. This suggests a broader principle: in any market with perfect substitution between altruistic and non-altruistic agents, costless reallocation undoes the altruists' actions, rendering their choices irrelevant for aggregate outcomes. Formalizing this as a general theorem—identifying the precise conditions on technology, preferences, and market structure under which altruistic behavior has zero impact—would unify the futility results across the demand side (Kaufmann et al., 2024), the investment side (Arnold, 2023), and the labor supply side studied here. The key conditions appear to be: (i) agents are perfect substitutes (no match-specific productivity), (ii) the relevant market clears frictionlessly, and (iii) the margin of altruistic action (firm choice, consumption, investment) does not affect the total quantity transacted. When any of these conditions fails—through match quality, search frictions, or elastic supply—futility breaks down.

- **Endogenous firm entry.** If firms choose their production technology (low vs. high externality) in response to labor supply conditions, altruistic sorting could affect the extensive margin of firm types. This connects to the literature on corporate social responsibility as a competitive strategy.

- **Leisure-labor tradeoff.** The baseline model assumes inelastic labor supply: each worker supplies one unit of labor regardless of the wage. Introducing a leisure-labor tradeoff (e.g., utility $v(c, \ell)$ where ℓ is leisure) would make labor supply upward-sloping in the wage. This has two important consequences. First, when altruists withdraw from the bad firm, normal workers no longer perfectly replace them—some normal workers respond to the lower wage at the bad firm by reducing their labor supply, so total employment at the bad firm falls. This breaks the futility result even when $M \leq 1/2$. Second, the analysis would become more closely analogous to the demand-side results of Kaufmann et al. (2024) and Trammell (2025): just as reducing consumption of an externality-generating good reduces its production, reducing labor supply to a high-externality firm would reduce its output and externality—but only partially, because general equilibrium feedback through wages dampens the effect. The key contrast between the supply and demand sides would then shift from complete futility versus partial impact to differences in the degree of equilibrium feedback.

- **Dynamic models.** In a multi-period setting, altruistic workers might build firm-specific human capital at clean firms, reducing the scope for GE substitution. Reputation effects and career concerns could also interact with altruistic motivation.

- **Naive altruistic workers.** The altruistic workers in the model are rational consequentialists: they correctly anticipate the equilibrium impact of their choices. An alternative is "naive" altruists who expect a one-to-one impact of their labor supply decisions on aggregate outcomes—for example, believing that by withdrawing from the bad firm they reduce pollution by one worker's worth, without accounting for the normal worker who replaces them. Kaufmann et al. (2024) study the analogous case on the consumption side and show that mixing naive and rational consumers can worsen outcomes relative to either group alone. A similar analysis on the labor supply side could yield a "naive sorting" equilibrium distinct from the Nash equilibrium studied here.

- **Warm-glow altruism.** The altruistic workers in this paper are consequentialists: they evaluate actions by their consequences for the externality E, not by the act itself. An alternative specification would feature "warm-glow" altruists (Andreoni, 1990) who derive utility from working at a clean firm per se, regardless of the equilibrium effect on the externality. Such workers would accept lower wages at the good firm even when general equilibrium substitution fully undoes the effect on pollution. This would break the futility result: warm-glow workers would sort into clean firms, generating wage differentials across firms and lowering their own consumption. Whether this sorting improves welfare depends on whether the experienced utility gain from working at the preferred firm outweighs the consumption loss—a question that requires specifying the source and magnitude of the warm glow.

- **Mixed consequentialist and deontological populations.** Sections 4.4 and 5.6 analyze the case where all altruists are deontologists. A natural extension is a mixed population: a fraction of altruists are consequentialists and the remainder are deontologists. Deontological workers would provide a rigid lower bound on clean-firm employment, while consequentialist workers adjust on the margin and exploit the earn-to-give channel. The equilibrium would feature richer sorting patterns than either pure case.

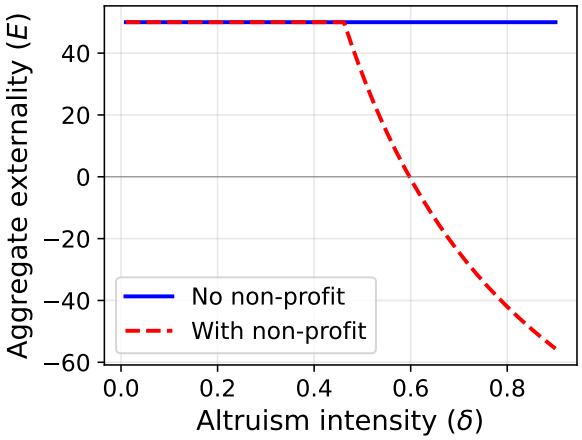
- **Public sector funded by taxation.** Consider adding a public sector of size L_{pub} determined by a policy decision and funded by a uniform tax on all workers. The public sector produces abatement (or other public goods) at rate ψ per worker, analogously to the non-profit. The key difference from the non-profit is that the public sector is large enough that altruists can never fill it entirely: $L_{\text{pub}} > ML$, so some normal workers always staff the public sector. Since the public sector must attract normal workers, it must pay the market wage w_n—the same wage normal workers earn in the private sector. This has a striking implication: altruists working in the public

sector generate *no marginal benefit* relative to normal workers, because each altruist who takes a public-sector job displaces a normal worker one-for-one, with no change in public-sector output or the externality. The futility logic from the private sector (Section **??**) reappears in a new guise. Since public-sector employment is capped by the policy-determined budget, the total number of public-sector workers is fixed regardless of who fills those positions. Altruists are therefore better off working in the private sector—either sorting into clean firms (reducing the externality through the compensating differential channel) or earning higher wages at dirty firms to fund non-profit donations. The only scenario in which an altruist might prefer the public sector is the earn-to-give channel: if the public sector happens to pay wages above an altruist's reservation consumption c_a^*, the altruist could work there and donate the surplus to the non-profit. But since the public sector pays w_n—the same wage available at normal-worker private firms—this offers no advantage over private-sector employment. In short, a tax-funded public sector that is large relative to the altruist population simply absorbs normal workers at the margin, and altruists have no comparative advantage in staffing it.

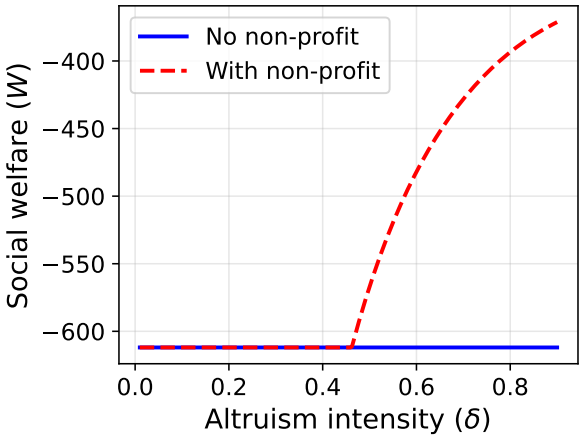
- **Warm-glow philanthropists.** A distinct application of warm-glow preferences concerns the donation channel rather than firm choice. In the baseline model, altruists donate because each dollar given to the non-profit reduces the externality by ψ/c_a—a purely consequentialist motive. If instead donors derive direct utility from the act of giving, as in Andreoni (1990), the donation first-order condition becomes $v'(c_a) = \delta\psi/c_a + \theta$, where $\theta > 0$ is a warm-glow parameter capturing the psychic benefit of donating. This has two implications. First, warm-glow donors contribute more than consequentialist donors at every level of δ, because the warm glow provides an additional motive beyond the externality reduction. Second, warm-glow donations remain positive even as the economy grows large and the individual impact on E becomes negligible—resolving the Andreoni (1988) concern that purely altruistic contributions shrink to zero. In the present model, the utilitarian microfoundation already ensures that consequentialist donations remain finite in large economies (Section 2.4), so the warm-glow is not needed for this purpose. However, warm-glow philanthropists would generate higher total donations and non-profit employment than the consequentialist baseline, and the welfare implications would depend on whether the warm-glow represents a genuine component of experienced utility or merely a behavioral bias.

Figure A.1: Non-profit sector (two-firm futility case)

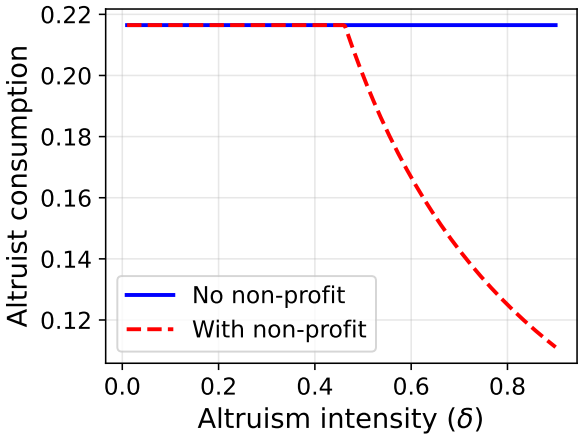
(a) Net externality



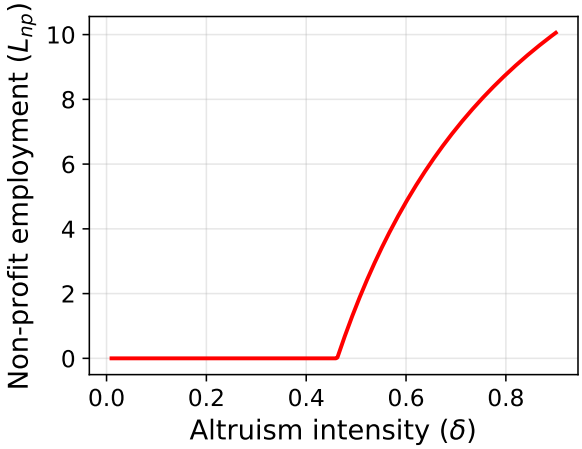
(b) Social welfare



(c) Altruistic worker consumption



(d) Non-profit employment



Notes: Equilibrium outcomes as altruism δ varies, for the two-firm special case ($J = 2$) where $M = 0.20 < 1/2$, so the futility result holds without a non-profit. Solid blue: without non-profit (flat, since sorting has no effect on outcomes). Dashed red: with non-profit ($\psi = 10.0$). Even when altruistic sorting is irrelevant, the non-profit provides an effective channel: donations fund abatement, reducing E and increasing welfare. δ ranges from 0 to $0.30\Gamma = 0.9$ (the range over which the approximate behavioral utility reliably represents the exact utilitarian objective; see Appendix D). Parameters: $A = 1.0$, $\alpha = 0.7$, $\phi = 1.0$, $\psi = 10.0$, $L = 100$, $M = 0.20$, $\gamma = 0.03$, $\sigma = 2$.

B Heterogeneous Productivity and Heterogeneous Altruism

The main text adopts two simplifying assumptions: common productivity across firms ($A_j = A$ for all j) and binary worker types (normal workers with $\delta = 0$ and altruistic workers with a common $\delta > 0$). This appendix sketches how the results extend when firms have heterogeneous productivity A_j and workers draw altruism δ_k from a continuous distribution F. The main results are:

1. **Heterogeneous productivity.** When firms differ in A_j, the no altruism equilibrium features unequal employment: $L_j^* = L \cdot A_j^{1/(1-\alpha)}/S$ where $S = \sum_j A_j^{1/(1-\alpha)}$. More productive firms employ more workers. The futility threshold becomes $M^* = A_{\min}^{1/(1-\alpha)}/S$, the employment share of the least productive firm.

2. **Heterogeneous altruism.** When workers have a continuous distribution of δ, the equilibrium features assortative matching: higher-δ workers sort into cleaner firms, with cutoff conditions determining which workers go where. The wage ladder reflects both the externality differential across firms and the marginal worker's altruism at each cutoff.

3. **Jensen's inequality and the distribution of altruism.** The donation function $d^*(\delta) = w^* - (\delta\psi)^{-1/(\sigma-1)}$ is strictly concave in δ for $w^* > c_a^*$. By Jensen's inequality, total donations are lower under a mean-preserving spread in the distribution of altruism. However, this intensive-margin effect can be offset by an extensive-margin effect: spreading the distribution activates donors in the right tail who would not donate under a degenerate distribution with low mean.

C Why Extreme Earn-to-Give Does Not Arise

A natural concern is whether the earn-to-give logic could push altruists toward the most harmful firms—taking a job at an outlier-high-externality employer to maximize donations. This appendix shows that, in the model's equilibrium, no wage premium is available at such firms, so the trade-off does not arise.

Proposition 13 (No Extreme Earn-to-Give Premium)**.** *Consider the many-firm model with J firms of common productivity A, ordered by externality rate $\phi_{(1)} < \cdots < \phi_{(J)}$. Suppose one firm—say firm J—has an outlier-large externality: $\phi_{(J)} \gg \phi_{(j)}$ for all $j < J$.*

1. ***Without sorting*** *($M \leq 1/J$): All firms pay the same wage $w^* = \alpha A(L/J)^{\alpha-1}$. Firm J pays w^* regardless of how large $\phi_{(J)}$ is. There is no wage premium at the harmful firm, so an altruist gains nothing by working there.*

2. ***With sorting*** *($M > 1/J$): In the sorting equilibrium (Proposition 6), normal workers equalize wages across all firms $m, m+1, \ldots, J$, including firm J. These firms all pay the common normal-firm wage w_n. The wage deviations occur only at the altruist-only firms $1, \ldots, m-1$, which pay* less *than w_n. Firm J pays exactly w_n—the same as every other normal-worker firm, regardless of the magnitude of $\phi_{(J)}$.*

3. ***With a non-profit:*** *The earn-to-give channel operates through the donation $d^* = w_j - c_a^*$. Since the marginal firm and all dirtier firms pay w_n, an altruist earns the same wage at firm J as at any other normal-worker firm. The donation is $d^* = w_n - c_a^*$, independent of the employer's externality rate.*

Proof. Parts (1) and (2) follow directly from the equilibrium characterization. Normal workers are indifferent across firms based solely on wages (they do not internalize the externality). With common productivity A, wage equalization requires equal employment at all firms where normal workers are present. In part (1), all firms employ L/J workers and pay w^*. In part (2), firms $m, \ldots, J$ each employ $(1-M)L/(J-m+1)$ normal workers (plus any altruists at the marginal firm m), all at wage $w_n = \alpha A((1-M)L/(J-m+1))^{\alpha-1}$. The externality rate $\phi_{(J)}$ does not enter any wage expression.

For part (3), the altruist's sorting condition with a non-profit (equation 27) determines wages at altruist firms as $w_j = w_n - c_a(\phi_{(m)} - \phi_{(j)})/\psi$. At any normal-worker firm (including firm J), the wage is w_n. The donation $w_n - c_a^*$ does not depend on which normal-worker firm the altruist chooses, so there is no incentive to select the most harmful one. □

The intuition is straightforward: because all workers (altruistic or not) are equally productive at every firm, and normal workers choose firms based solely on wages, normal workers flood any firm that would otherwise offer a premium—including the most harmful firm. Competitive pressure from normal workers bids the wage at firm J down to the common level w_n. The only firms that deviate from w_n are the clean firms, where altruists willingly accept a wage penalty. An outlier-large externality at one firm has no effect on that firm's wage, because the externality enters neither the production function nor the normal workers' objective.

This result means that in the model, an altruist never faces the dilemma of choosing between a low-paid ethical job and a high-paid unethical one: the "unethical" firms pay the same as the mid-range firms. The earn-to-give trade-off is between the clean firms (which

pay below w_n) and the normal-wage firms (which pay w_n), not between ethical employment and a lucrative but harmful outlier. The extreme philosophical scenario—"should I work at the worst company to fund the most donations?"—does not arise as an equilibrium outcome.

Heterogeneous altruism and the outlier firm

The equal-productivity baseline above shows that no wage premium exists at harmful firms. A natural extension asks: what if all workers are at least slightly altruistic, so that the most harmful firm is understaffed—and therefore *does* pay a premium? Even in this case, it is not the most altruistic workers who fill those positions.

Consider J firms with a continuous distribution of altruism δ_k across workers, where $\delta_k \geq \underline{\delta} > 0$ for all k (everyone is at least slightly altruistic). If $\underline{\delta}$ is large enough that no worker with $\delta = \underline{\delta}$ is willing to work at firm J at the equal-employment wage, then firm J is understaffed in equilibrium. With fewer workers, its marginal product is higher, so it pays a premium $w_J > w_n$.

This premium is real—it reflects the scarcity of workers willing to tolerate the extreme externality. One can think of financial fraud or other ethically dubious activities as analogous: they may pay well precisely because few people are willing to engage in them.

However, the sorting condition implies that the workers who *do* accept employment at firm J are those with the *lowest* δ. The behavioral utility of working at firm J is $v(w_J) - \delta\phi_J + \delta\psi d_J/c_a$ (accounting for the earn-to-give channel). A worker with high δ faces a large behavioral penalty $\delta\phi_J$ from the extreme externality. For this to be offset by the wage premium and donation potential, the premium $w_J - w_n$ would need to scale with δ—but it does not, because w_J is determined by the marginal (least altruistic) worker at firm J, not by the most altruistic potential worker.

Formally, in the sorting equilibrium with heterogeneous δ, worker k weakly prefers firm j over firm j' if and only if

$$v(w_j) - \delta_k\phi_j + \delta_k\frac{\psi(w_j - c_a^*(\delta_k))}{c_a^*(\delta_k)} \geq v(w_{j'}) - \delta_k\phi_{j'} + \delta_k\frac{\psi(w_{j'} - c_a^*(\delta_k))}{c_a^*(\delta_k)}.$$

Since the donation term $\psi(w_j - c_a)/c_a$ is the same at any two firms paying the same wage, and w_J is only modestly above w_n (determined by the marginal worker with δ close to $\underline{\delta}$), a worker with $\delta \gg \underline{\delta}$ strictly prefers any cleaner firm: the externality penalty $\delta(\phi_J - \phi_j)$ grows linearly in δ, while the wage premium $w_J - w_n$ does not.

The upshot: even when the most harmful firm pays a premium, the model predicts that it is staffed by the *least* altruistic workers, not the most altruistic. A "hardcore" altruist

with high δ would lose more in behavioral utility from the extreme externality than they would gain from the wage premium and subsequent donations. The earn-to-give logic does not push the most dedicated altruists toward the most harmful activities—it pushes the least altruistic workers there, because they face the smallest behavioral cost from the externality.

D Quality of the Behavioral Utility Approximation

The main text uses the behavioral utility $v(c) - \delta E$ as the altruist's objective, where $\delta \equiv \lambda\Gamma$ and $\Gamma = L\gamma$. This approximation, derived in Section 2.4, drops a term capturing how each worker's actions affect other workers' consumption through general equilibrium wage adjustments. This appendix examines when the approximation works well and when it generates counterintuitive results.

The exact utilitarian utility. A worker with altruism weight λ who cares about all workers' experienced utility has the exact objective

$$u_i^{\text{exact}} = v(c_i) + \lambda \sum_{k \neq i} \big[v(c_k) - \gamma E\big] - \gamma E. \tag{30}$$

The main-text approximation replaces (30) with $v(c_i) - \delta E$, dropping the consumption terms $\lambda \sum_{k \neq i} v(c_k)$ on the grounds that any individual worker's action has a negligible effect on other workers' consumption. This is correct for each individual: the effect of one worker's donation on the market wage w_n is $O(1/L)$, which vanishes as $L \to \infty$. However, when M altruists simultaneously donate, the collective effect on w_n is $O(M/L)$, which does not vanish. The approximation captures the externality channel (each dollar donated reduces E by ψ/c_a) but misses the wage channel (donations collectively reduce private-sector employment, raising w_n for remaining workers).

When the approximation works well. For the sorting equilibrium without a non-profit (Section 4), the approximation is accurate because altruists only choose which firm to work at—they do not withdraw labor from the private sector. The wage channel is absent, and the behavioral utility $v(c_i) - \delta E$ correctly captures the altruist's trade-off between consumption and externality reduction.

For the non-profit equilibrium at moderate δ, the approximation is also adequate. When donations are modest, the collective effect on w_n is small, and the approximate and exact equilibria are close.

When the approximation breaks down. At high δ (highly altruistic workers), the approximate model produces a counterintuitive result: social welfare *declines* with the non-profit active, even though the non-profit reduces the externality. The mechanism: the approximate first-order condition $v'(c_a) = \delta\psi/c_a$ gives $c_a = (\delta\psi)^{-1/(\sigma-1)}$, which approaches zero as δ grows. With CRRA utility ($\sigma > 1$), the welfare cost of near-zero consumption is severe, and in the approximate model this cost exceeds the externality benefit.

This welfare decline does not occur in the exact model. Figure D.1 compares the two. In the exact model, each altruist accounts for the positive wage externality of donations (higher w_n for normal workers), which partially substitutes for the externality-reduction motive and leads to higher c_a (less extreme self-sacrifice). Formally, for symmetric altruists the exact equilibrium c_a maximizes

$$u(c_a;\lambda) = \big(1+\lambda(M-1)\big)\, v(c_a) + \lambda N \cdot v\big(w_n(c_a)\big) - \big[\gamma + \lambda(L-1)\gamma\big] \cdot E(c_a), \tag{31}$$

where $w_n(c_a)$ and $E(c_a)$ are determined by the equilibrium at that c_a.[14] In the exact model, welfare is monotonically increasing in λ: more altruism always weakly improves social welfare. At $\lambda = 1$ (perfect utilitarianism), the exact Nash equilibrium coincides with the social planner's optimum, as it should—the game is a potential game with potential W.

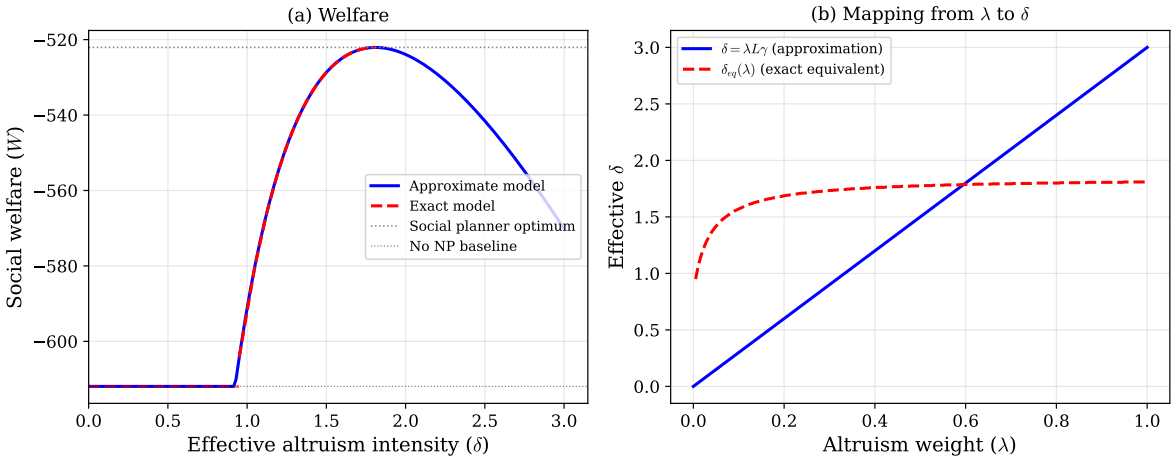
Implications for the main results. The main-text simulations with the non-profit (Figures 8 and 9) restrict δ to moderate values where the approximation is accurate. The sorting results (Sections 3–4) are unaffected by the approximation, as the wage channel is absent without a non-profit. The qualitative insights of the non-profit analysis—that the donation channel weakens sorting, that consequentialist altruists are indifferent across strategies, and that deontological altruists generate less abatement—hold in both the approximate and exact models.

E Alternative Title Suggestions

- Effective Altruism at Work: Altruistic Labor in General Equilibrium
- Is Altruism Futile? Altruistic Labor in General Equilibrium
- Does Altruism Work? Altruistic Labor in General Equilibrium
- Altruistic Labor in General Equilibrium

[14] The term $\lambda N \cdot v(w_n(c_a))$ captures the wage channel: as c_a decreases (more donation), L_{np} increases, L_{priv} decreases, and w_n rises, benefiting normal workers.

Figure D.1: Exact vs. approximate behavioral utility



Notes: Social welfare and altruist consumption as the altruism weight λ varies from 0 (no altruism) to 1 (perfect utilitarianism), for the two-firm model with a non-profit. Blue solid: approximate behavioral utility $v(c) - \delta E$. Red dashed: exact utilitarian utility (31). Gray dotted: social planner's optimum (first-best). The approximate model produces a welfare decline at high λ because altruists over-donate (panel b), driving consumption too low. The exact model avoids this: altruists internalize the wage benefit of their donations, donate less aggressively, and welfare increases monotonically. Parameters: $A = 1.0$, $\alpha = 0.7$, $\phi = 1.0$, $\psi = 5.0$, $L = 100$, $M = 20$, $\gamma = 0.03$, $\sigma = 2$.