A theory of good intentions[*]

Paul Niehaus
UC San Diego & NBER

January 1, 2020

Abstract

Altruists' effectiveness often falls short of their intentions. To better understand this, I model an altruist who derives warm glow from his perceptions of outcomes, as opposed to the outcomes themselves. This altruist rationally avoids some (but not all) information that could inform his actions, and all feedback on their effects. Intermediaries such as charities can increase revenue by limiting the supply of specific kinds of information, depending on donor motives. Beneficiaries (or regulators) may also prefer to limit information provision as they manage a tradeoff between the quantity and the quality of giving.

[*] I thank Nageeb Ali, Jim Andreoni, Roland Bénabou, Navin Kartik, Lant Pritchett, Joel Sobel, Adam Szeidl, Lise Vesterlund, and participants in various seminars for their helpful comments. Microsoft Research New England provided generous hospitality during early work on this project. While this paper was written with the best of intentions, it predicts that it will have less impact than I had hoped.

1 Introduction

Altruists often seem to produce results that fall short of their intentions – so much so that the term "well-intentioned" has become a euphemism for "poorly informed." For example, Americans give a generous 2% of GDP to charity each year, yet only 3% of these donors even *claim* to have done any research comparing the effectiveness of alternatives.[1] This begs the question: if people really are well-intentioned, why don't they *become* well-informed?

One possibility is that altruists want to learn, but find it costly or difficult. For small charitable donations, the costs of learning can exceed the benefits (Krasteva and Yildirim, 2013). Market failures limit the supply of information, which may be a public good (Duflo and Kremer, 2003; Levine, 2006; Ravallion, 2009). And the creation and communication of information by practitioners can be distorted by strategic considerations (Pritchett, 2002; Duflo and Kremer, 2003; Levine, 2006). These observations have inspired a number of reforms, including the creation of modern research institutions such as J-PAL, IPA, and CEGA.

This paper examines a second, complementary possibility: altruists do not want to achieve a better outcome, but rather to *believe* they have done so. This creates tension, as perception and reality can diverge. For example, consider sponsoring a child in a developing country: a donor may experience a "warm glow" (Andreoni, 1989) *thinking* that the child is eating well and attending school, even if in reality he has gone hungry or flunked out. This begs the question how learning works in a market where such thoughts and perceptions *are* the "product." Specifically, the paper considers a single altruist (abstracting from issues of public goods) with arbitrary other-regarding preferences. It then builds a model of altruistic behavior from three basic ingredients. First, the altruist may not learn the results of his actions. Second, the altruist tends to interpret the resulting uncertainty optimistically, to maximize his psychological well-being. And third, there are limits to this optimism; the altruist cannot entirely self-delude.[2] It then examines in turn how these assumptions shape the preferences over information structures of various players in the market: the altruist himself, intermediaries seeking to maximize revenue, and the beneficiary.

The altruist avoids all feedback about the effects his actions have had, preferring to assume that no news is good news. He avoids some information that could inform his decisions, but not all: he prefers to learn ex ante anything that he must eventually learn ex post. If a donor knows that results from an impact evaluation will eventually be released, for example, he would prefer to see them before giving, to guard against disappointment.

Revenue-maximizing intermediaries (e.g. charities) tailor their strategies to the specifics of the altruist's other-regarding preferences. Broadly speaking, they should ignore wishful beliefs that motivate giving while confronting with data those that do not. If the altruist wants to maximize his perceived impact (Duncan, 2004), for example, then the charity should suppress all information; the altruist wishes to believe that the return to his giving is high, and the

[1] Giving statistics: author's calculation using data from The Giving Institute (2013) and the Bureau of Economic Analysis (`http://www.bea.gov/national/index.htm#gdp`, accessed 7 August 2013). Research statistics: see Hope Consulting (2012). The Hope sample over-represents wealthier donors and thus if anything likely overestimates the research done by an average donor.

[2] The model thus builds on evidence that people tend to interpret ambiguous (and even unambiguous) information in self-serving ways (e.g. Eil and Rao (2011), Mobius et al. (2013)).

charity's best strategy is not to interfere. If the altruist's preferences are "pure" then the charity should suppress information about effectiveness (which the altruist wishes to believe) but generate information about the beneficiary's need (which the altruist wishes to disbelieve). This is consistent with nonprofit marketing strategies that emphasize "awareness-raising" and graphic depictions of need ("poverty pornography") over evidence of cost-effectiveness, and with the overall scarcity of information about charitable impact.[3]

Beneficiaries (or third parties) have mixed incentives to generate information. One the one hand, information may sometimes counteract the altruist's tendency to become over-enthusiastic about relatively ineffective causes. Yet for the same reason information tends to make the altruist less enthusiastic about giving overall. The beneficiary may thus faces a tradeoff between the *quality* and the *quantity* of giving. There are times, consequently, when it is best not to look a gift horse in the mouth.

The model is consistent with some recent laboratory evidence. Several studies find that subjects prefer to avoid ex-post feedback on how their actions affected others (Dana et al., 2007; Fong and Oberholzer-Gee, 2011; Grossman and van der Weele, forthcoming), and one finds that the *anticipation* of ex-post feedback reduces their ex ante efforts to learn (Jhunjhunwala, 2017). This latter result in particular suggests a close connection between donors' naivete and the absence of direct feedback they receive when helping others. Recent developments in real-world charitable giving mirror this, as practitioners have grown skeptical of donors' interest in learning. The Hewlett Foundation recently ended a $12M initiative to promote evidence-based giving, for example, because "the initiative assumed that donors would use this information if they could find it... [but] most donors aren't even looking."[4] More broadly, the model may speak to phenomena such as the rapid adoption of new approaches to foreign aid that capture the imagination of practitioners and grow into a large industry before any rigorous evidence on their impact is available (Banerjee et al., 2015),[5] or the (low) quality of personal gift-giving (Waldfogel, 2009).[6]

Conceptually the paper builds on three lines of work. First, it takes seriously Andreoni's (1989) influential idea that altruists derive "warm glow" from their actions. As Andreoni et al. (2017) have emphasized, "the warm-glow hypothesis simply provides a direction for research rather than an answer to the puzzle of why people give – the concept of warm-glow is a placeholder for more specific models of individual and social motivations." This paper offers one such model, linking warm glow to perceived outcomes.

Second, it builds on work on motivated reasoning (Bénabou and Tirole, 2016). If, as this literature has argued, people prefer to hold positive beliefs about their own abilities, futures,

[3] Charity evaluator GiveWell writes, for example, that "useful information about what different charities do and whether it works isn't publicly available anywhere." `http://www.givewell.org/about/story`, accessed 10 September 2013.

[4] Video interview with Lucy Bernholz, `http://www.hewlett.org/programs/effective-philanthropy-group`, accessed 18 May 2014.

[5] Easterly (2006) emphasizes the role played by faith and desire in such phenomena: "I feel like kind of a Scrooge... I speak to many audiences of good-hearted believers in the power of Big Western Plans to help the poor, *and I would so much like to believe them myself*" [emphasis added].

[6] Unwanted Christmas gifts are common enough that there are websites devoted to displaying examples: see `www.badgiftemporium.com` or `whydidyoubuymethat.com`.

etc., then it seems plausible that they also prefer to hold positive beliefs about their impacts on others.[7] Indeed, in the model the altruist endogenously holds biased beliefs only about the likelihood of events he never observes; in this sense the paper studies a relatively modest departure from fully rational expectations compared to others, and in particular the optimal expectations framework of Brunnermeier and Parker (2005), on which it builds. The paper differs otherwise primarily in its application, focusing on characterizing the demand for and supply of information in the marketplace for doing good.

Third, it adds to the work above asking why useful information about how to do good is so limited. Pritchett (2002) argues that "this dearth of knowledge is sufficiently striking as to deserve explanation and common explanations casually proposed – ethical barriers, costs, and feasibility – are not sufficient." He focuses on strategic interactions between program advocates and potential funders, while I build from frictions generated by the preferences of the funders themselves.

The paper proceeds as follows: Section 2 presents the framework and characterizes optimal beliefs, Section 3 derives preferences over information structures, and Section 4 describes outstanding questions.

2 The framework

2.1 Timing and uncertainty

There are two players, an altruist and a beneficiary. Play evolves as follows:

1. Nature selects a finite-valued parameter $\theta \in \Theta$
2. The altruist observes a signal $s_1 \in S_1$ and forms subjective ex ante beliefs $\hat{\pi}_1(\theta, s_2|s_1)$
3. The altruist chooses a decision $d \in D$
4. The altruist observes a signal $s_2 \in S_2$ and forms subjective ex post beliefs $\hat{\pi}_2(\theta|s_2, s_1)$
5. Payoffs are realized

Let $\pi(\theta, s_1, s_2)$ be the true joint distribution of the unobservable parameter θ and the observable data (s_1, s_2). Expectations are with respective to π except where otherwise noted.

2.2 Payoffs

The beneficiary's payoff $w(d, \theta)$ depends on the decision d and state θ. The altruist's payoff depends on two components: a purely self-interested component $u(d)$, and an other-regarding component $v(d, \theta)$ which represents the utility the altruist obtains from the beneficiary's outcome(s). If the altruist were perfectly informed about these outcomes, his payoff would be

$$u(d) + v(d, \theta) \tag{1a}$$

[7] See for example Akerlof and Dickens (1982), Caplin and Leahy (2001), and Bénabou and Tirole (2002), among many others.

When this is not the case his payoff must instead depend on his perception of θ:

$$u(d) + \mathbb{E}_{\hat{\pi}_2}[v(d, \theta)] \tag{1b}$$

This captures the idea that uncertainty about θ may not resolve completely by the time payoffs are realized.

Equation (1b) models altruism quite generally, in the sense that the beneficiary's preferences $w(\cdot, \cdot)$ and the altruist's preferences for her $v(\cdot, \cdot)$ can be arbitrarily related to each other. They could even capture misanthropy, e.g. $v = -w$. Below I provide some general results but also examine in more detail special cases of interest suggested by the literature. Under *pure altruism* $v = w$, so that – conditional on the level of u – the altruist and the beneficiary agree on how to assess the beneficiary's well-being. Alternatively, the altruist might have paternalistic preferences. Duncan (2004) has argued that some altruists care about their *impact*, which corresponds here to the difference between the beneficiary's payoff and the counterfactual payoff they would have realized without the altruist's help. Defining $\overline{d} = \arg\max_d u(d)$ as the counterfactual action, an impact philanthropist is defined by

$$v(d, \theta) = w(d, \theta) - w(\overline{d}, \theta) \tag{2}$$

Andreoni et al. (2017) argue that guilt must be an important motivator for some givers as many of them "avoid the ask" to give. Interpreting guilt as a desire to reduce the gap between what one does and what one *should* do, we can represent it using (2) with the reference action $\overline{d}$ defined as the one that should be taken.

2.3 Optimization

The altruist maximizes subjective expected utility by choosing a decision given his beliefs

$$d^*(\hat{\pi}_1, s_1) \in \arg\max_d u(d) + \mathbb{E}_{\hat{\pi}_1}[v(d, \theta)] \tag{3}$$

and beliefs given a decision rule

$$(\hat{\pi}_1^*, \hat{\pi}_2^*) \in \arg\max_{\hat{\pi}_1, \hat{\pi}_2} \mathbb{E}\Big[u(d^*(\hat{\pi}_1, S_1)) + \mathbb{E}_{\hat{\pi}_2}[v(d^*(\hat{\pi}_1, S_1), \theta)]\Big] \tag{4}$$

Beliefs play distinct roles in each period: ex ante beliefs determine the altruists' decision, while ex post beliefs determine how he interprets its consequences. These beliefs need not be consistent with Bayes' rule, though it will turn out that optimal beliefs are. I require only that beliefs be consistent with the true data generating process in the following sense:

Assumption 1 (Admissible beliefs)**.** *Subjective beliefs satisfy* $\hat{\pi}_1(\theta, s_2|s_1) = 0$ *if* $\pi(\theta, s_2|s_1) = 0$ *and* $\hat{\pi}_2(\theta|s_1, s_2) = 0$ *if* $\pi_2(\theta|s_1, s_2) = 0$.

This imposes a degree of logical consistency: the altruist understands that some events are impossible and cannot wishfully believe otherwise. If he could believe anything at all then he would always hold the rosiest possible views; the more interesting case is the one in which there is some cost to holding increasingly implausible beliefs. Assumption 1 is one simple way of capturing that idea.

2.4 Optimal beliefs

Having made a decision and observed signals, the altruist chooses to be as optimistic as possible about the consequences, obtaining payoff $u(d) + \overline{v}(d, s_1, s_2)$ where

$$\overline{v}(d, s_1, s_2) \equiv \max_{\theta \in \Theta(s_1, s_2)} [v(d, \theta)] \tag{5}$$

and $\Theta(s_1, s_2) \equiv \{\theta : \pi(\theta|s_1, s_2) > 0\}$ contains the support of any admissible belief. It will be convenient henceforth to work with $\overline{v}$, which summarizes how information affects the altruist's payoffs. Optimal ex ante beliefs solve

$$\max_{\hat{\pi}_1} u(d^*(\hat{\pi}_1, s_1)) + \mathbb{E}\left[\overline{v}(d^*(\hat{\pi}_1, s_1), s_1, S_2)\right]] \tag{6}$$

This is a standard choice under uncertainty problem, except that the uncertainty that matters is about S_2 rather than θ. This is because the altruist cares not about the outcome per se, but about the constraints on what he will be able to believe about the outcome, which are determined by S_2. It is in his interest to hold Bayesian beliefs about those constraints. Specifically, let $a(s_1) : S_1 \rightarrow D$ be any decision rule and let $f(\theta|a(s_1), s_1, s_2)$ be any conditional distribution function that assigns positive probability only to elements of $\Theta(s_1, s_2)$ that solve (5) when $d = a(s_1)$. Then beliefs derived as conditional probabilities from the prior

$$\hat{\pi}(\theta, s_1, s_2) = f(\theta|a(s_1), s_1, s_2)\pi(s_1, s_2) \tag{7}$$

are optimal given $a(s_1)$, and uniquely so up to equivalencies. One way to think of this is that the altruist holds an unbiased view $\pi(s_1, s_2)$ of the likelihood of the various signals he will observe, but interprets those signals in an optimistic way.

Lemma 1 (Baysian Updating)**.** *There exist optimal beliefs of the form defined by applying Bayes' Rule to (7), and any beliefs that do not induce a payoff distribution identical to some such belief are not optimal.*

Proof. It is straightforward to check existence by applying Bayes' Rule to (7) with $a(\cdot)$ any decision rule that is optimal given beliefs. To show the converse, consider arbitrary beliefs $\tilde{\pi}$ that do not "conform" to the definition about and that induce action profile $\tilde{a} : S_1 \rightarrow D$. If this action profile can be generated by some other conforming beliefs π, then the payoff distributions induced by $\tilde{\pi}$ and π must differ due only to differences in the posterior beliefs; but since by (5) conforming beliefs yield optimal posterior beliefs, this means the payoff distribution generated by $\tilde{\pi}$ must be dominated weakly everywhere and strictly in at least one case by that generated by π. Suppose alternatively that the action profile $\tilde{a}$ differs from that induced by any conforming beliefs, so that in particular $\tilde{a}(s_1) \neq a(s_1)$ for some realization s_1 and some action profile a induced by confirming beliefs π. If $(\tilde{a}, \tilde{\pi})$ yield a strictly higher expected payoff than (a, π) conditional on s_1 then π cannot conform to the definition above, while if they yield the same expected payoff then a modified version of a which takes action $\tilde{a}(s_1)$ must also be optimal given beliefs π, again contradicting the premise. □

This has several implications. First, optimal beliefs are consistent with observable data: the marginal distribution over (s_1, s_2) implied by (7) is the empirical distribution $\pi(s_1, s_2)$. The difference between optimal beliefs and reality lies only in the areas in which reality is itself unobservable (i.e. the conditional distribution of θ given (s_2, s_1).). The beliefs of an altruist who learns the model through repeated experience could thus converge to optimal beliefs. In this sense the model represents a modest departure from fully rational expectations relative to other models of motivated reasoning. Consequently, the model's empirical content relative to rational expectations lies not in beliefs about and realizations of the observable data (S_1, S_2) but rather in how manipulating the information structure affects beliefs and behavior, as for example in the experimental design of Jhunjhunwala (2017).

Second, optimal beliefs are self-consistent: an altruist holding them would not wish to alter them. Mathematically, $\hat{\pi}$ is a fixed point of (7), though the empirical distribution π need not be. This property typically does not hold in models with tension between utility from actions and from beliefs such as Brunnermeier and Parker (2005).

Third, the model nests the benchmark case of preferences over outcomes. If the feedback (s_1, s_2) that the altruist receives always uniquely identifies the state θ then it is "as if" θ were observed ex post, just as in the benchmark case.[8] This underscores that the model's interesting properties result from ex-post uncertainty about θ. While the core idea here is that this kind of uncertainty often arises when helping other people, since the decision-maker does not personally experience the results, it could also arise in the context of self-regarding decisions such as whether to purchase a credence good.

3 Good intentions and learning

I next endogenize the information structure, examining in turn the preferred information structures of the altruist, an intermediary, and the beneficiary himself. The relative importance of these preferences varies by context: whether and how the parties can communicate, market structure, and so on. Intermediaries such as charities, for example, often decide what information to generate about their work and how accessible to make it to donors. If charities compete by committing to provide donors with the information they want, however, donor preferences will still carry weight. Beneficiaries have little influence over what altruists learn in some cases (e.g. a farmer who receives a cow paid for by a foreign donor), while in other cases they have more (e.g. a friend or family member sharing information about the gifts they would like to receive).

I focus throughout on the decision to generate an informative, publicly observed signal. This might correspond for example to deciding whether or not to commission a randomized controlled trial to evaluate the impact of an intervention. Another useful exercise, which I

[8] Formally, call realizations (s_1, s_2) revealing if $\Theta(s_1, s_2)$ has a single element and call signals (S_1, S_2) fully revealing if any realization is revealing. Then $\overline{v}(d, s_1, s_2) = v(d, \theta)$ for any realization (s_1, s_2) and (6) reduces to the standard problem of choice under uncertainty about θ.

do not pursue here, would be to study communication in games with asymmetric information in which the intermediary or beneficiary observes private signals and sends messages to the donor.

3.1 The altruist

As is well known, a decision-maker who cares about outcomes places weakly positive value on information ex ante and none ex post, while one who places intrinsic value in his beliefs may value information negatively (Golman et al., 2017). Here the costs of information are particularly direct, as it tightens the plausibility constraint (Assumption 1). I next examine the patterns of avoidance and demand this generates. I use standard terminology: random variables X and Y are *informationally equivalent* if there exists a bijection f such that $Y = f(X)$; Y is more informative than X with respect to Z given joint distribution $h(x, y, z)$ if $h(x|y, z)$ is independent of z.

To build intuition, consider an altruist who has sponsored a child overseas. If he receives no feedback, he can choose to believe that the child is doing well in school, generating warm glow. If on the other hand he receives information about the results there is some chance he will learn the child is struggling in school, forcing him to revise his optimistic views downward. Thus, the value of ex post feedback is unambiguously negative.

Proposition 1. *The altruist's expected payoff is weakly lower the more informative is ex post feedback.*

Proof. A sketch of the argument is as follows: let random variable S_2' be a garbling of S_2 with respect to (S_1, θ). Even fixing the altruist's decision at the value d^* he would choose when he expects to see S_2, he must do at least weakly better when observing S_2' since it is a garbling of S_2 and thus constrains his ex-post beliefs less tightly. Thus he must certainly do better when also allowed to re-optimize his choice of d. Details are in the online appendix. □

The comparison will be strict whenever there is some chance that the additional information will rule out a state θ that maximizes $v(d, \theta)$ for some decision d the altruist chooses with positive probability.

Ex ante, before the altruist makes his decision, things are more complicated. Consider learning whether or not a sponsored child has a chance of graduating – whether the school he attends is a good one, for example. Per se, this again constrains what the altruist can believe and is thus unappealing. But now suppose he anticipates that he will *eventually* learn whether or not the child graduates. Because this information will directly affect his payoff, he values the ability to forecast it and make decisions accordingly. More generally, he values ex ante information to the extent it lets him forecast ex post information, but no further.

Proposition 2. *Fix the information the altruist receives ex post.*

(a) The altruist's expected payoff is weakly greater when he observes equivalent information ex ante than when observing any other ex ante information.

(b) If the altruist's ex ante information is less (more) informative than his ex post information, then his expected payoff is weakly increasing (decreasing) in the informativeness of the former.

Proof. (a) Consider a signal S_1 that is equivalent to S_2. The argument is as follows: the altruist cannot do better than if he could choose a decision d after observing *only* S_2, since (i) the set of tenable beliefs after observing both S_1 and S_2 is weakly smaller than after observing only S_2, and (ii) S_1 obviously contains no information more useful for predicting S_2 than does S_2 itself. A formal statement is in the appendix.

(b) Let S_1 garble S_2 with respect to θ and let S_1' garble S_1 with respect to S_2. I make the standard argument that information weakly improves decision-making, with the wrinkle that we must also show that it does not in this case tighten the constraint on beliefs in period 2. Formally, fix a realization s_1 of S_1. The altruist's expected payoff if he observes S_1' is

$$\begin{aligned}\mathbb{E}\left[\max_d u(d) + \mathbb{E}\left[\max_{\theta\in\Theta(S_2,S_1')} v(d,\theta)|S_1'\right]|s_1\right] &= \mathbb{E}\left[\max_d u(d) + \mathbb{E}\left[\max_{\theta\in\Theta(S_2,s_1)} v(d,\theta)|S_1'\right]|s_1\right] \\ &\leq \mathbb{E}\left[\max_d u(d) + \mathbb{E}\left[\max_{\theta\in\Theta(S_2,s_1)} v(d,\theta)|s_1\right]|s_1\right] \\ &= \max_d u(d) + \mathbb{E}\left[\max_{\theta\in\Theta(S_2,s_1)} v(d,\theta)|s_1\right]\end{aligned}$$

which is his expected payoff if he observes S_1. The first equality holds because $\Theta(S_1,S_2) = \Theta(S_1',S_2) = \Theta(S_2)$ since both S_1 and S_1' are garblings of S_2 with respect to θ, while the second inequality holds because s_1 predicts S_2 more accurately than S_1'.

Now consider a S_1 that is a refinement of S_2 with respect to θ and let S_1' be a refinement of S_1 with respect to S_2. Since both ex ante signals are finer than S_2, the set of ex post admissible beliefs will be independent of S_2 in either case. If the altruist observes S_1 his payoff will thus be

$$\max_d \left[u(d) + \max_{\theta\in\Theta(S_1)} v(d,\theta)\right] \leq \max_d \left[u(d) + \max_{\theta\in\Theta(S_1')} v(d,\theta)\right]$$

which is his payoff if he observes S_1', where the inequality follows from the fact that $\Theta(S_1') \subseteq \Theta(S_1)$ if S_1' is a refinement of S_1.

□

One corollary is that the value of observing a signal equivalent to S_2 is (weakly) positive, since it must be weakly greater that the value of receiving no information. Overall, the altruist's first motive is to avoid information, which limits what he can plausibly believe. This motive dominates ex post. Ex ante, however, the altruist is also motivated to gather information that will help him predict the (informational) constraints he can expect to face ex post, which generates a secondary, positive demand for information.

3.2 Intermediaries

To examine which strategies maximize revenue for an intermediary such as a charity, interpret d as the (real-valued) amount of money the altruist donates. To obtain interpretable comparative statics, let Θ be ordered and $v(d, \theta)$ be increasing in both arguments. Let

$$e(\mathcal{I}_1, \mathcal{I}_2) \equiv \mathbb{E}\left[\arg\max_d u(d) + \mathbb{E}\left[\overline{v}(d, \mathcal{I}_2)|\mathcal{I}_1\right]\right] \tag{8}$$

denote the donor's expected donation when he uses information $\mathcal{I}_1$ ex ante to form his forecast of ex post signals, and is constrained by information $\mathcal{I}_2$ ex post when forming his beliefs.

Mathematically, the key consideration for the charity is whether states of the world that the altruist views as good ones (i.e. that yield high $v(d, \theta)$) tend to make him more or less likely to give. If states θ that yield a high v are also associated with high *marginal* returns v_d to donations, then the altruist wants to believe that returns are high. He will will thus convince *himself* to give generously, without any outside intervention. In this case the charity should simply keep quiet. If, on the other hand, states that yield a high v are associated with low marginal returns, then the altruist will convince himself that the returns to giving are low. In this case the charity can profitably intervene.

Proposition 3. *Expected revenue is lower (higher) when ex post information is more informative if v is supermodular (submodular)*

Proof. Let S_2 be a refinement of S_2' with respect to θ. Conditional on s_1, we can write the altruists objective function as

$$h(d, \{x(s_1, s_2', s_2)\}) \equiv u(d) + \sum_{s_2} \sum_{s_2'} v(d, x(s_1, s_2', s_2))\pi(s_2'|s_2)\pi(s_2|s_1)$$

where $x(s_1, s_2', s_2) = \max\{\theta : \pi(\theta, s_1, s_2) > 0\}$ $(\max\{\theta : \pi(\theta, s_1, s_2') > 0\})$ if he observes S_2 (S_2'). Note that we can write the distribution of S_2' in this separable form because it garbles S_2, and that x does not depend on d since v is monotone in θ. Examining f, its latter argument is an element of a lattice with dimension support$(S_2) \times$ support(S_2'); moreover since S_2' garbles S_2 we have $\max\{\theta : \pi(\theta, s_1, s_2') > 0\} \geq \max\{\theta : \pi(\theta, s_2, s_1) > 0\}$ for any realization (s_2', s_2), so that S_2' induces a weakly larger element of this lattice than S_2. It then follows from the monotone comparative statics theorem of Milgrom and Shannon (1994) that the solution is weakly greater (smaller) under S_2' if v is supermodular (submodular). □

Ex ante information, on the other hand, has two distinct effects. First, it constrains the altruist's beliefs, just as ex post information does. This effect tends to generate the same comparative statics as above. But it also enables the altruist to better predict ex post information, as in a standard model. This effect tends to generate ambiguous comparative statics: generically, more information may either raise or lower the expected action. Formally, we can decompose the expected revenue effects of revealing a signal S_1 rather than some S_1'

which garbles it as

$$\underbrace{e(\{S_1\},\{S_1,S_2\})-e(\{S_1\},\{S_1',S_2\})}_{Constraining}+\underbrace{e(\{S_1\},\{S_1',S_2\})-e(\{S_1'\},\{S_1',S_2\})}_{Predicting} \tag{9}$$

where the first term captures the constraining effect and the second the predictive effect.

Proposition 4. *More informative ex ante information decreases (increases) expected revenue through the constraining effect in (9) if v is supermodular (submodular). Moreover, it has the same effects on overall expected revenue if (u,v) respect expectation, in the sense that*

$$\arg\max_d \mathbb{E}_\mu[u(d)+v(d,\theta)] = \mathbb{E}_\mu[\arg\max_d u(d)+v(d,\theta)]$$

for any $\mu \in \Delta(\theta)$

Proof. Let S_1' be a garbling of S_1 with respect to (S_2,θ).

(a) It is enough to show the result for any particular realization (s_1,s_1'). Consider therefore

$$\arg\max_d u(d)+\mathbb{E}\left[\overline{v}(d,s_1,S_2)|s_1\right]-\arg\max_d u(d)+\mathbb{E}\left[\overline{v}(d,s_1',S_2)|s_1\right]$$

By the argument used in proving Proposition 3 this difference is negative (positive) if v is supermodular (submodular).

(b) Fix a realization s_1' of S_1' and for this section let expectations refer to expectations conditional on this realization. The prediction effect is

$$\mathbb{E}\left[\arg\max_d u(d)+\mathbb{E}\left[\overline{v}(d,s_1',S_2)|S_1\right]\right]-\arg\max_d u(d)+\mathbb{E}\left[\overline{v}(d,s_1',S_2)\right] \tag{10}$$

Define $\overline{\theta}=\arg\max_{\theta\in\Theta(s_1',S_2)} v(d,\theta)$ and let $\overline{\pi}(\overline{\theta},s_1,s_2)$ be the joint distribution of $(\overline{\theta},s_1,s_2)$. Then we can write (10) as

$$\mathbb{E}\left[\arg\max_d u(d)+\mathbb{E}_{\overline{\pi}}\left[v(d,\overline{\theta})|S_1\right]\right]-\arg\max_d u(d)+\mathbb{E}_{\overline{\pi}}\left[v(d,\overline{\theta})\right]$$

Since preferences respect expectation we can write this as

$$\mathbb{E}\left[\mathbb{E}_{\overline{\pi}}\left[\arg\max_d u(d)+v(d,\overline{\theta})|S_1\right]\right]-\mathbb{E}_{\overline{\pi}}\left[\arg\max_d u(d)+v(d,\overline{\theta})\right]$$

Since π and $\overline{\pi}$ agree on the marginal distribution of S_1, this is zero by the law of iterated expectations.

□

The condition in (b) says that preferences are such that information would have no generic tendency to increase or decrease giving in a standard model with $S_2=\theta$.[9] [10] Notice that the

[9] An example of preferences that respect expectation is $u(d)=y-\frac{d^2}{2}$ and $v(d,\theta)=d\theta$.

[10] The additive separability of (9) suggests that if this condition holds approximately then part (b) should also

constraining effect is zero in the benchmark case where the donor learns the truth ex post (i.e. $S_2 = \theta$) since in this case S_1 contains no incremental information to further constrain beliefs.

3.2.1 Marketing strategy

How should a charity market itself in light of the above results? The answer turns out to depend on the relationship between the preferences of the altruist (v) and those of the beneficiary (w).

Consider first pure altruism, i.e. $v = w$. In this case the charity should suppress information about its own effectiveness. Information about effectiveness means information about a parameter θ that is complementary to d in w (and hence in v), and so Proposition 4 implies the charity should suppress it. Intuitively, the donor already prefers to believe that his gift is highly impactful, and the charity can only make things worse by providing information that may contradict that belief. The charity should provide information, however, about the neediness of the beneficiary. Suppose for example that θ measures the beneficiaries other sources of income, and hence substitutes for d in w (and hence in v). The altruist will prefer to believe that θ is high and therefore that the marginal value of his giving is low. The charity can motivate more giving by forcing him to accept that the beneficiary is in fact in need of help. This distinction may help explain some common patterns in nonprofit marketing, such as the tendency to emphasize "awareness-raising" and information about need (e.g. by depicting forlorn children or starving famine victims) rather than concrete information about what will be done with donations and how effective it is.

Now consider an altruist motivated by impact (Duncan, 2004). He wants to believe that he is "making a difference," i.e. that the return v_d to giving is high. He therefore wishes to believe both that the charity he funds is effective, and also much needed. Since both of these beliefs stimulate giving and thus benefit the intermediary, its best strategy is to provide no information, leaving the altruist as much space as possible for wishful thinking.[11]

Finally, consider altruists motivated by guilt. A guilty giver presents the opposite challenge: to minimize his guilt he prefers to believe that the need is small and that intervention would be ineffective anyway, rationalizing a small donation d. To raise money, the charity should seek to alter both of these beliefs.

Formally, let preferences be as in (2) and $w_d(d, \theta)$ be monotone in θ. If the reference action is the least generous one ($\overline{d} = \min D$) I say that the altruist is motivated purely by impact, while if it is the most generous one ($\overline{d} = \max D$) he is motivated purely by guilt.

Proposition 5. *Expected revenue decreases (increases) in the informativeness of ex post feedback if the altruist is motivate by impact (guilt). Moreover, the same holds for the informativeness of ex ante information if (u, v) respect expectation.*

Proof. (a) Let S_2' be a garbling of S_2 with respect to θ. Given d and the realization (s_1, s_2)

hold approximately.

[11] This case may also help explain the success of "matching grant" campaigns, where a lead donor promises to give only if others do as well. While promises like these might seem non-credible, an impact donor has a strong incentive to believe them as they imply that his own gift will have substantial "leverage."

the altruist's ex-post problem is

$$\max_{\theta \in \Theta(s_2, s_1)} w(d, \theta) - w(\overline{d}, \theta)$$

Since $w_d(d, \theta)$ is monotone in θ, the solution to this problem must also solve $\max_{\theta \in \Theta(s_2,s_1)} w_d(d, \theta)$ for *any* d if $d \geq \overline{d} = \min D$, and $\min_{\theta \in \Theta(s_2,s_1)} w_d(d, \theta)$ for *any* d if $d \leq \overline{d} = \max D$. It follows that further constraining the altruist's ex-post beliefs by revealing additional information will decrease (increase) the expected value of $v_d(d, \theta)$ for any d, and thus weakly decrease (increase) his expected donation, when $\overline{d} = \min D$ ($\overline{d} = \max D$).

(b) Let S_1' be a garbling of S_1 with respect to (S_2, θ). The argument proceeds exactly as in the proof of Part 2 of Proposition 4. The effect of coarser information has two effects, a constraint effect and a prediction effect; the prediction effect is zero when preferences respect expectation, while the sign of the constraint effect depends on $\overline{d}$ as in part (a) above.

□

Overall, the best marketing strategy for an intermediary depends on the specific motivations of the donors it faces. Charities should segment their audiences and their communication according to these motivations, or if this is difficult may specialize in raising money from a particular type. All else equal, communication to altruists motivated by guilt should be most informative, while (ironically) communication to those motivated by impact should be least informative. More broadly, the model suggests that increasing a gift's salience or memorability should increase giving by increasing the weight the altruist places on his thoughts $\mathbb{E}[v]$. This is one way to interpret common nonprofit marketing practices such as sending "thank-you" notes, or offering "gift catalogues" that afford donors little real control but let them visualize their gifts as providing some specific, memorable item.

3.3 Beneficiaries

To understand the beneficiary's incentives to generate information, it is helpful to first consider how her preferences $w(d, \theta)$ differ from those of the altruist (1b). First, the altruist may be paternalistic, wanting different things for the beneficiary than she wants for herself (i.e. $v \neq w$). Second, the beneficiary generally wants the altruist to be more generous, in the sense of accepting a lower value of $u(d)$, since she puts higher relative weight on her own payoff w than the altruist does. These two wedges are common in models of altruism. The third, which is not, is that the beneficiary cares about her actual payoffs while the altruist cares about his perceptions of those payoffs. If the altruist donates an animal, for example, the beneficiary cares whether or not the animal actually gets sick and dies, while the altruist cares about whether he can reasonably *believe* that the animal is alive. Focusing on this third tension, we can draw out two novel implications for learning. First, more information can reduce the expected *quality* of giving; and second, even when it increases quality, this may come at a cost in terms of the *quantity* of giving.

To see why quality may fall, consider a pure altruist ($v = w$) choosing whether to spend a fixed amount of money buying animals (a) or bednets (b) for a beneficiary. The (uncertain) values of each gift to the beneficiary are

$$\theta_a \in \{VH, M\}$$
$$\theta_b \in \{H, L\}$$

where $VH > H > M > L$. In other words, the value of animals may be very high, but may also be less than that of bednets. As long as there is *some* chance that the returns to animals is very high (VH) the altruist will give animals and believe that it is, even if bednets provide better *expected* value from the beneficiary's point of view (e.g. if $\theta_a = M$ and $\theta_b = H$ with high probability). The problem is that the beneficiary cares about the expected actual outcome, while the altruist cares about the expected best plausible outcome.

Generating information can help or hurt the beneficiary in this situation. Consider for example the effects of a signal that reveals whether or not $\theta_a = VH$. If $\theta_b = H$ and $\theta_a = M$ with very high probability, this signal almost surely helps the beneficiary: if it reveals $\theta_a = VH$ then the altruist will continue to donate to a, while if it reveals $\theta_a \neq VH$ he will switch to b which has higher conditional expected value. But if $\theta_b = L$ and $\theta_a = M$ with very high probability, this signal may hurt the beneficiary: if it reveals $\theta_a \neq VH$ the altruist will switch to b and convince himself that $\theta_b = H$, when in fact the beneficiary would have preferred that he stuck with a. While obviously stylized, this illustrates the idea that information can make the beneficiary worse off in expectation because she expects to disagree with the altruist about how to interpret it.

Even when information improves the quality of the altruist's giving, this may come at the cost of quantity, as the wishful thinking that leads an altruist to back *relatively* ineffective causes also induces him to be *absolutely* more generous. For example, he may become excited about a novel and relatively untested approach to alleviating poverty precisely because it has not yet been proven *not* to have very high returns, while neglecting older approaches with proven, modest returns. To illustrate this, consider an altruist allocating money between donations $\{d_a, d_b\}$ to causes $\{a, b\}$ as well as own consumption $y - d_a - d_b$, with preferences

$$\begin{aligned} u(d) &= u(y - d_a - d_b) \\ v(d,\theta) &= w(d,\theta) = \theta_a d_a + \theta_b d_b \end{aligned} \tag{11}$$

He expects to receive no feedback (i.e. $S_2 = \emptyset$). A charity evaluator can conduct a (costless) ex ante experiment to reveal the return parameters $\{\theta_c\}$. If it does not, the altruist will believe the best that is *plausible* about both causes, i.e. will believe that $\theta_c = \overline{\theta}_c \equiv \max\{\theta_c | \mathbb{P}(\theta_c) > 0\}$, and give an amount an amount $d^*(\overline{\theta}_c^*)$ defined by $u'(y - d^*(\overline{\theta}_{c^*})) = \overline{\theta}_{c^*}$ to the cause c^* with the highest plausible return. The beneficiary's payoff is $\theta_{c^*} d^*(\overline{\theta}_{c^*})$.

Now suppose an evaluation reveals the return θ_c to each cause. The donor gives an amount $d^*(\theta_{c^{**}})$ to the best cause $c^{**} = \arg\max_c \theta_c$, with $d^*(\cdot)$ as defined above. The beneficiary's payoff is $\theta_{c^{**}} d^*(\theta_{c^{**}})$. Comparing this to her previous payoff and taking expectations, the welfare effects of revealing θ can be decomposed as

$$\underbrace{\mathbb{E}[(\theta_{c^{**}} - \theta_{c^*})d^*(\theta_{c^{**}})]}_{\text{Accuracy}} + \underbrace{\mathbb{E}[\theta_{c^*}(d^*(\theta_{c^{**}}) - d^*(\overline{\theta}_{c^*}))]}_{\text{Discouragement}}$$

The first term captures the benefit of more effective giving, holding its *level* fixed; it is weakly positive, and strictly so provided there is some positive probability that information will change the donors' choice of cause c. The second term, on the other hand, captures the cost of disillusioning the donor: because reality tends to be worse than he had hoped, he tends to give less generously when forced to confront it.

This logic may help to explain why people who *know* that the altruistic projects they work on are ineffective nevertheless hesitate to speak up about this. They are not, in this view, cynically seeking to protect their own rents. Rather, they correctly assess that raising concerns will depress support for work in their area more generally, and could thus ultimately do more harm than good.

4 Conclusion

Models of other-regarding behavior typically specify preferences over outcomes, abstracting from the fact that the decision-maker may never experience these outcomes. This paper has examined whether and when this distinction matters, studying an altruist with preferences over his beliefs about the beneficiary's outcomes. The model nests the standard one in the special case where ex post feedback is complete, but differs otherwise. In particular, the altruist endogenously prefers to avoid ex post feedback entirely and to learn ex ante only enough to predict ex post feedback; revenue-maximizing intermediaries such as charities suppress specific kinds of information depending on donor motives; and beneficiaries may suppress information expected to reduce either the quality or the quantity of giving.

While static, the framework is dynamically consistent in the sense that the altruist's beliefs match the empirical distribution of observable variables. Modelling a dynamic extension could potentially shed further light on the evolution of altruism. Two specific conjectures may merit examination. First, altruistic behavior self-perpetuates. An altruist who takes an arbitrary action will be motivated to believe in the future that this action was effective, which will in turn motivate him to repeat it. Initial donor acquisition will thus be crucial for charities. Second, altruists may become "jaded" over time as the accumulation of evidence increasingly limits their ability to "think positive."

References

Akerlof, George A and William T Dickens, "The Economic Consequences of Cognitive Dissonance," *American Economic Review*, June 1982, *72* (3), 307–19.

Andreoni, James, "Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence," *Journal of Political Economy*, December 1989, *97* (6), 1447–58.

_ , Justin M. Rao, and Hannah Trachtman, "Avoiding the Ask: A Field Experiment on Altruism, Empathy, and Charitable Giving," *Journal of Political Economy*, 2017, *125* (3), 625–653.

Banerjee, Abhijit, Esther Duflo, Rachel Glennerster, and Cynthia Kinnan, "The Miracle of Microfinance? Evidence from a Randomized Evaluation," *American Economic Journal: Applied Economics*, 2015, *7* (1), 22–53.

Bénabou, Roland and Jean Tirole, "Self-Confidence and Personal Motivation*," *The Quarterly Journal of Economics*, 08 2002, *117* (3), 871–915.

_ and _ , "Mindful Economics: The Production, Consumption, and Value of Beliefs," *Journal of Economic Perspectives*, September 2016, *30* (3), 141–64.

Brunnermeier, Markus K. and Jonathan A. Parker, "Optimal Expectations," *American Economic Review*, September 2005, *95* (4), 1092–1118.

Caplin, Andrew and John Leahy, "Psychological Expected Utility Theory And Anticipatory Feelings," *The Quarterly Journal of Economics*, February 2001, *116* (1), 55–79.

Dana, Jason, Roberto Weber, and Jason Kuang, "Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness," *Economic Theory*, October 2007, *33* (1), 67–80.

Duflo, Esther and Michael Kremer, "Use of randomization in the evaluation of development effectiveness," Technical Report, World Bank 2003.

Duncan, Brian, "A theory of impact philanthropy," *Journal of Public Economics*, August 2004, *88* (9-10), 2159–2180.

Easterly, Bill, *The White Man's Burden: Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good*, Oxford University Press, 2006.

Eil, David and Justin M. Rao, "The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself," *American Economic Journal: Microeconomics*, 2011, *3* (2), 114–38.

Fong, Christina and Felix Oberholzer-Gee, "Truth in giving: Experimental evidence on the welfare effects of informed giving to the poor," *Journal of Public Economics*, 2011, *95* (5), 436–444.

Golman, Russell, David Hagmann, and George Loewenstein, "Information Avoidance," *Journal of Economic Literature*, March 2017, *55* (1), 96–135.

Grossman, Zachary and Joël van der Weele, "Self-Image and Strategic Ignorance in Moral Dilemmas," *Journal of the European Economic Association*, forthcoming.

Hope Consulting, "Money for Good: The US Market for Impact Investments and Charitable Gifts from Individual Donors and Investors," Technical Report, Hope Consulting May 2012.

Jhunjhunwala, Tanushree, "Feedback, Search and Charitable Giving," mimeo, Department of Economics, The Ohio State University 2017.

Krasteva, Silvana and Huseyin Yildirim, "(Un)Informed Charitable Giving," *Journal of Public Economics*, 2013, *106*, 14–26.

Levine, David, "Learning What Works – and What Doesn't: Building Learning into the Global Aid Industry," Technical Report, UC Berkeley 2006.

Milgrom, Paul and Chris Shannon, "Monotone Comparative Statics," *Econometrica*, January 1994, *62* (1), 157–80.

Mobius, Markus, Muriel Niederle, Paul Niehaus, and Tanya Rosenblat, "Managing Self-Confidence: Theory and Experimental Evidence," Technical Report, UC San Diego November 2013.

Pritchett, Lant, "It pays to be ignorant: A simple political economy of rigorous program evaluation," *Journal of Policy Reform*, 2002, *5* (4), 251–269.

Ravallion, Martin, "Evaluation in the Practice of Development," *World Bank Research Observer*, March 2009, *24* (1), 29–53.

The Giving Institute, *Giving USA 2013*, Giving USA Foundation, 2013.

Waldfogel, Joel, *Scroogenomics: Why You Shouldn't Buy Presents for the Holidays*, Princeton University Press, 2009.

A More detailed proofs

Proof of Proposition 1

Let random variable S_2' be a garbling of S_2 with respect to (S_1, θ). Fix a realization s_1 and let d^* be an action that is optimal given this realization if the donor expects to then observe the realization of S_2. If instead he observes the realization of S_2' he can still choose d^*, and hence his payoff cannot be less than

$$u(d^*) + \mathbb{E}\left[\max_{\theta\in\Theta(s_1,S_2')} v(d^*,\theta)|s_1\right] = u(d^*) + \mathbb{E}\left[\mathbb{E}\left[\max_{\theta\in\Theta(s_1,S_2')} v(d^*,\theta)|s_1, S_2\right]|s_1\right] \tag{12}$$

$$\geq u(d^*) + \mathbb{E}\left[\mathbb{E}\left[\max_{\theta\in\Theta(s_1,S_2)} v(d^*,\theta)|s_1, S_2\right]|s_1\right] \tag{13}$$

$$= u(d^*) + \mathbb{E}\left[\max_{\theta\in\Theta(s_1,S_2)} v(d^*,\theta)|s_1\right] \tag{14}$$

which is his payoff when observing S_2. The first equality holds by the law of iterated expectations. The inequality holds because $\Theta(s_1, s_2) \subseteq \Theta(s_1, s_2')$; to see this, note that if θ is possible given (s_1, s_2) (i.e. $\pi(s_1, s_2, \theta) > 0$) then for any realization s_2' such that $\pi(s_2'|s_2) > 0$, θ must also be possible given (s_1, s_2'), as the fact that S_2' garbles S_2 implies we can write $\pi(\theta, s_1, s_2') = \pi(s_2'|s_2)\pi(\theta, s_1, s_2) > 0$.

Proof of Proposition 2

Proof. (a) Consider a signal S_1 that is equivalent to S_2. His payoff is

$$\mathbb{E}\left[\max_d u(d) + \mathbb{E}\left[\max_{\theta\in\Theta(S_2,S_1)} v(d,\theta)|S_1\right]\right] \leq \mathbb{E}\left[\max_d u(d) + \mathbb{E}\left[\max_{\theta\in\Theta(S_2)} v(d,\theta)|S_1\right]\right] \tag{15}$$

$$\leq \mathbb{E}\left[\mathbb{E}\left[\max_{d,\theta\in\Theta(S_2)} u(d) + v(d,\theta)|S_1\right]\right] \tag{16}$$

$$= \mathbb{E}\left[\max_{d,\theta\in\Theta(S_2)} u(d) + v(d,\theta)\right] \tag{17}$$

which is his payoff if $S_1 = S_2$. The first inequality holds since the set of tenable beliefs after observing both S_1 and S_2 is weakly smaller than after observing only S_2, and the second holds since the altruist cannot do worse if he can condition is choice of d on the realization of S_2.

□