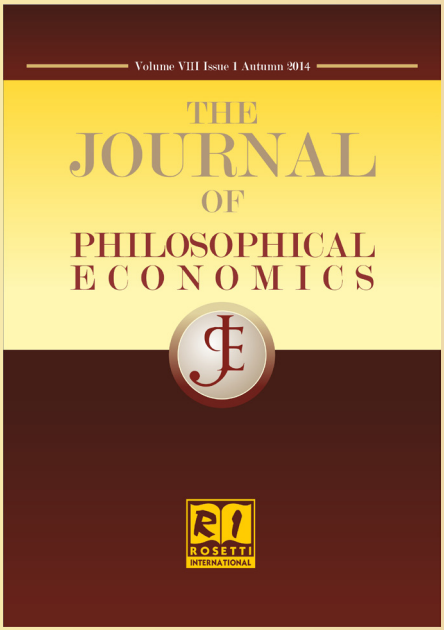
Modeling exogenous moral norms

Ross A Tippit

THE JOURNAL OF
PHILOSOPHICAL ECONOMICS

Volume VIII Issue 1 Autumn 2014

ISSN 1843-2298



Modeling exogenous moral norms

Ross A. Tippit

ROSETTI
INTERNATIONAL

Tippit, Ross A. (2014) 'Modeling exogenous moral norms',
The Journal of Philosophical Economics, VIII:1

Modeling exogenous moral norms

Ross A. Tippit

Abstract: This paper considers the possibility of a robust and general formulation of a model of choice for the representation of a variety of moral norms. It starts by reviewing several recent models of deontological (or rule-based) norms that retain the basic elements of the economic model of choice. It briefly examines the achievements and drawbacks of each model, and while no model is identified as the most accurate or robust, the most appealing aspects of each model contribute to the construction of a *tout-ensemble* utility function proposed in the final section. This representation of preferences aims to incorporate the most common qualities of both consequentialist and deontological moral norms in order to represent decision making under their influence.

Keywords: morals, norms, deontic, choice, decision, nonconsequentialist

Introduction

Motivation

What started in the latter half of the 19th century as a model of consumer choice has expanded to a model for explaining a broad range of human conduct and behavior. While the influence of moral norms on economic activity has been noted by economists since the inception of the discipline, positing moral beliefs as exogenous, independent impetuses of human activity has proved challenging. In the last two decades economists have increasingly approached moral norms— whether they are stipulated as a form of reciprocity, altruism, or fairness— as exogenous elements of choice for use in particular research topics. In like manner, moral norms have increasingly been approached as endogenous variables in economic models [1]. Since the mid-1980s it was clear that moral norms of a specific variety posed a theoretical

Received : 24 April 2014

difficulty for the model of choice [2]. The inherent consequentialist orientation of the model of choice presented a challenge for modeling moral beliefs of a non-consequentialist or deontological nature [3]. By the 1990s, enough literature had accumulated on the topic for reviews (of the literature) to take stock of the issue (see for example, Goldfarb and Griffith 1991a, 1991b; Koford and Miller 1991; Broome 1992).

Recently, several attempts have been made to represent moral norms of a deontological type within a model of choice. In some cases, these models were employed for a specific research topic, and in others, the models simply attempt to model agency under the influence of moral norms. This paper reviews some of these recent attempts, notably the work of Timur Kuran (1998), Joseph Heath (2008), Mathew Rabin (1995), Eyal Zamir and Barak Medina (2010), David C. Rose (2011), as well as Richard Dowell, Robert Goldfarb, and William Griffith (1998).

Following the review of each model in the next section, the strengths and insights of each are incorporated into a single model of choice. Aspiring to capture both consequentialist and deontic moral norms, this *tout-ensemble* model has a greater range of applicability than the models reviewed, and provides the basis for further theoretical development and experimental research.

Methodological concerns

The methodological and conceptual difficulties inherent in the incorporation of non-consequentialist evaluations of an agent's opportunity set are numerous and have been widely noted. The methodological difficulties vary from the philosophical and conceptual [4] to the mathematical and logistic [5]. This paper will not address or attempt to resolve these difficulties, but instead it will be assumed that our intuitions regarding deontological moral norms can be gainfully approached with the basic elements of the theory of choice, and that such modeling might help decipher the perimeters of any pitfalls.

The economist is tempted to translate moral norms into one of the three elements of choice: (1) a choice variable (element of the agent's objective function), (2) the constraint, or (3) the agent's preferences. They thereby leave themselves with three (not necessarily exclusive) possible approaches to modeling moral norms. This, in fact, might provide a brief account of economists' approaches to the challenge in the literature thus far [6].

The first of these three routes would render moral norms endogenous, and it might seem obvious that deciding upon which of the remaining two approaches to employ (and to what extent) depends on the type of moral norm one is attempting to model. For example, a deontic or categorical moral judgment would appear— with its division of actions into categories of permissible, forbidden, etc.— to make the opportunity set discontinuous. Thus, as constraints in the model of choice divide the potential option set into feasible and non-feasible— and as moral rules 'feel like' constraints on action— it seems logical to depict such moral norms as constraints of a sort. One might ask whether this 'discontinuity and restraint' is better achieved through a formulation of preferences. In fact, this choice may be arbitrary, so long as what is represented is a partition of the opportunity set and an inability to stipulate 'rates of substitution' between morally incomparable activities. We will explore this possibility through the course of the paper.

While adaptability and applicability are of great value to economic models, saving the appearance of choice through different models does not make the models epistemologically equivalent. There may be subtle differences (between constraints and preferences, for example) that have important implications for the meaning of the moral norm being modeled, even though the same choice can be represented by a variety of models. A model that bears realism in its depiction is not only intellectually compelling, but likely to be closer to the vehicle of causation. A guiding principle hereinafter is that a model should attempt to capture the nature of a moral norm in its effect as well as in its sense of importance in our choices. Thus, value will also be given to the proximity of a model to our intuitions regarding moral beliefs and values. We will now consider specific attempts at exogenously modeling deontic moral norms.

Recent models of moral norms

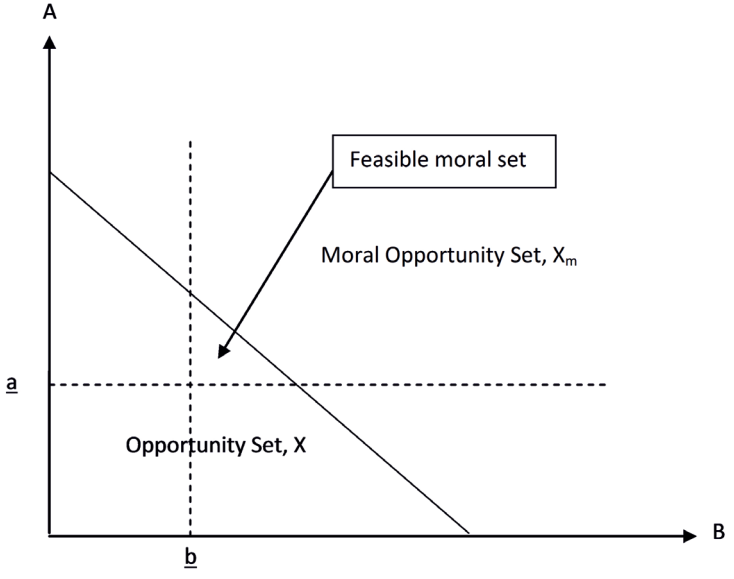
Breakable constraints— Kuran (1998)

Timur Kuran aims to understand the phenomenon of 'a guilty conscience' in the jargon of economics, highlighting the point that the 'bad feeling' is not 'fear of punishment,' but instead a feeling of not being satisfied with the preferences one's choice has revealed. This bad feeling emerges because 'your choice failed to accommodate a value that you hold' (Kuran 1998, p.231). In effect, then, a conscience imposes disutility when a choice satisfies preferences that your values do not agree with[7]. Kuran defines 'moral overload' as a condition where, given

physical and financial constraints, an individual's set of values cannot be satisfied (Kuran 1998, p.233). In the process of examining this condition, he gives a fairly simple, yet intuitively powerful formalization of moral norms as constraints that can be broken.

An agent's options are described by the opportunity set, *X*, which is 'determined jointly by his physical and economic constraints'; the agent also has 'internalized values that require him to avoid a subset of his options. Of the remaining possibilities, those within *X* form his *moral opportunity set*, X_m. If X_m contains at least one element, the underlying morality is *feasible*' (Kuran 1998, p.234).

Figure 1



In Figure 1, the area under the economic constraint represents the agent's opportunity set, *X*, the set of possible combinations of activities *A* and *B* under the constraint. The agent's moral norms prescribe that $a \geq \underline{a}$ and $b \geq \underline{b}$. Kuran calls the combination $(\underline{a}, \underline{b})$ the 'moral base' and is represented by $x_m = (\underline{a}, \underline{b})$, while an agent's actual choice is denoted by $x_c = (a, b)$.

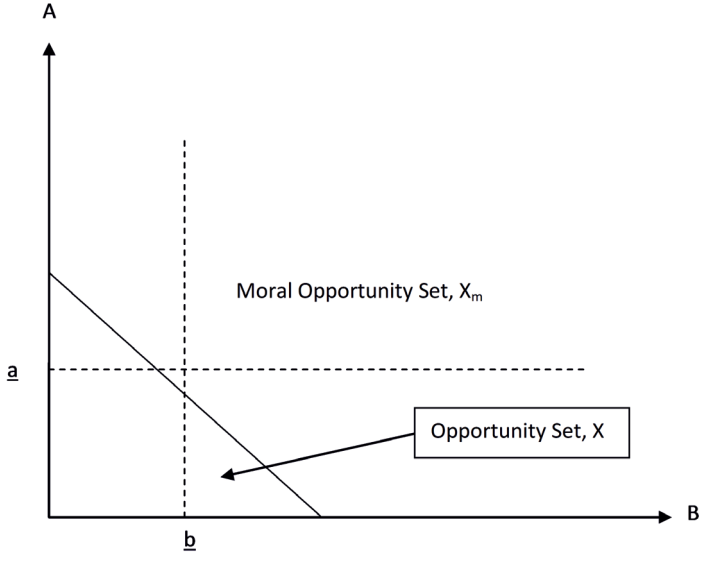
A subset of the opportunity set in Figure 1 is the intersection of the moral opportunity set, X_m, and the opportunity set, *X*. When X_m is nonempty, the morality

is feasible. However, a person's moral system may be infeasible; for example, the same values (as in Figure 1) may be imposed on a smaller opportunity set, and render an empty moral opportunity set. Thus, infeasibility

> can be caused by a shrinking of X, [...] or a hardening of the individual's values, or some combination of these changes. Whatever the cause of the infeasibility, the result is moral overload: The individual cannot possibly live by his values (Kuran 1998, p.234).

Choosing x_m would leave the individual 'at ease,' even though they might 'derive greater satisfaction from choices for which $a \leq \underline{a}$, $b \leq \underline{b}$, or both' (Kuran 1998, pp.234-35). A choice outside X_m may produce a feeling of guilt— or as Kuran calls it, 'moral dissonance.' The magnitude of the dissonance may depend on the distance between x_m and the individual's actual choice x_c. Kuran proposes that the actual choice be determined by the agent's preferences and constraints. Thus, we can imagine a case where x_c necessarily lies outside X_m, as depicted in Figure 2. In this case, 'moral dissonance is inescapable' (Kuran 1998, p.236). As Kuran defines it, then, moral dissonance can occur even if the agent cannot possibly meet their moral base, x_m.

Figure 2



Since the choice actually made, x_c, may result in disutility for the agent (due to the moral base or moral obligations not being met), Kuran proposes that the agent's 'total utility' be the sum of 'intrinsic utility,' $I(x)$, and 'moral utility,' $M(x, x_m)$ (Kuran 1998, p.243). Thus, total utility is:

$$U(x) = I(x) + M(x, x_m). \qquad (1)$$

He proposes that one's moral utility $M(x, x_m)$ be negative if X_m is empty and be zero if x_m is met (there does not appear to be any reason why satisfying the moral base does not render a positive contribution to utility). Moral dissonance, however, is equivalent to a negative value of moral utility. Kuran proposes that

> at least in the short run, the individual's only decision variable is x. Accordingly, he selects x to maximize $U(x)$, subject to his budget constraint and his predetermined morality. His choice x_c may well produce moral dissonance (Kuran 1998, p.243).

The agent's preferences are ultimately an 'ordering of all possible x, as measured by total utility, $U(x)$. This ordering takes account of any guilt that might be involved. But guilt is not decisive to the choice' (Kuran 1998, p.243). In other words,

> of two options that generate different amounts of guilt, the one that produces more might provide greater total utility. Hence, the possibility of conflict between preferences and values even with a morality that is feasible (Kuran 1998, p.243).

Kuran's model succinctly depicts moral norms as constraints that are breakable, and captures an intuition regarding our preferences over material affairs and their relation to moral norms. The agent may 'avoid making a morally satisfying selection even when able to do so' (Kuran 1998, p.236). According to Equation 1, a conflict between preferences and moral norms is by no means necessary: it is possible that preferences ($I(x)$) instruct a more generous level of either element beyond the moral base ($\underline{a} < a$, $\underline{b} < b$). The agent's preferences may induce a choice 'beyond one's obligation,' so long as the morality is feasible[8]. That said, one feature of moral norms as constraints is their 'rigidity' or resistance to trade-offs.

It is often the case that moral values express themselves as obligations or 'absolute commitments,' a case in which they resist formulation as preferences[9]. Trade-offs are usually represented through preferences; whereas, as Kuran puts it, moral norms render 'judgments that are more or less independent of possibilities' (Kuran 1998, p.236). Preferences simply rank one possible option more than another. Of course, a formulation can be given for $M(x, x_m)$ that is compatible with Kuran's two assumptions (when it takes a negative value and when it takes a value of zero),

in which case the effect of a morality ultimately shows itself as part of an agent's preferences. As such, the moral constraint acts through the agent's preferences, even though it may still operate to 'forbid' an action. This implication of modeling moral norms through preferences will reemerge later when the important qualities of each model are used to construct a general formulation.

Deontic trimming— Heath (2008)

A characteristic of deontic moral norms is that actions are evaluated and predicated of in a categorical manner; that is, the most general categorization of actions is into 'permissible' and 'forbidden.' As noted above, these divisions are rigid to the point that, if a ranking of actions exists, it is discontinuous in nature, suggesting that actions cannot be substituted with each other according to their moral value. A branch of logic, i.e. deontic logic, analyzes the relationships between these categories, and their implications for normative arguments and reasoning. Kuran (1998) proposed dividing an opportunity set into the moral opportunity set and its compliment; one interpretation of this is that the opportunity set is divided according to the two deontic categories of permissible and forbidden. The approach taken by Joseph Heath is less accommodating to the 'breakability' of moral constraints.

Heath's work spans many topics of philosophy and economics, and is generally concerned with the role of social norms in solving the coordination problem of 'social cooperation.' According to Heath, economic explanations of choice struggle with the necessity of rules. They are necessary directives as to what should be done; 'rules proscribe actions, not outcomes' (Heath 2008, p.66). By themselves, they do not provide incentives. In his view, the problem economics has with explaining rule-following 'stems from a tension between the motivational psychology of rule-following and the instrumental conception of rationality' (Heath 2008, p.66) [10], and the framework for analyzing decision making in economics is inherently inept at handling moral norms. As Heath explains: 'Rules usually classify actions as permissible or impermissible; they do not specify which outcomes are more or less desirable' (Heath 2008, p.66). He therefore suggests expanding the vocabulary of the economist for rule-following choices in a very straightforward way:

> Desires are intentional states associated with outcomes, beliefs are intentional states associated with states. Norms, or rule-following considerations more generally, can be accommodated simply by positing an intentional state associated directly with actions (Heath 2008, p.72).

For Heath, 'it is natural to think of desires as a set of preferences over outcomes, and principles as a set of preferences over actions,' since it 'formally parallels the existing structures of beliefs and desires— taking principles as given, in the same way the decision theorists take preferences as given' (Heath 2008, p.72).

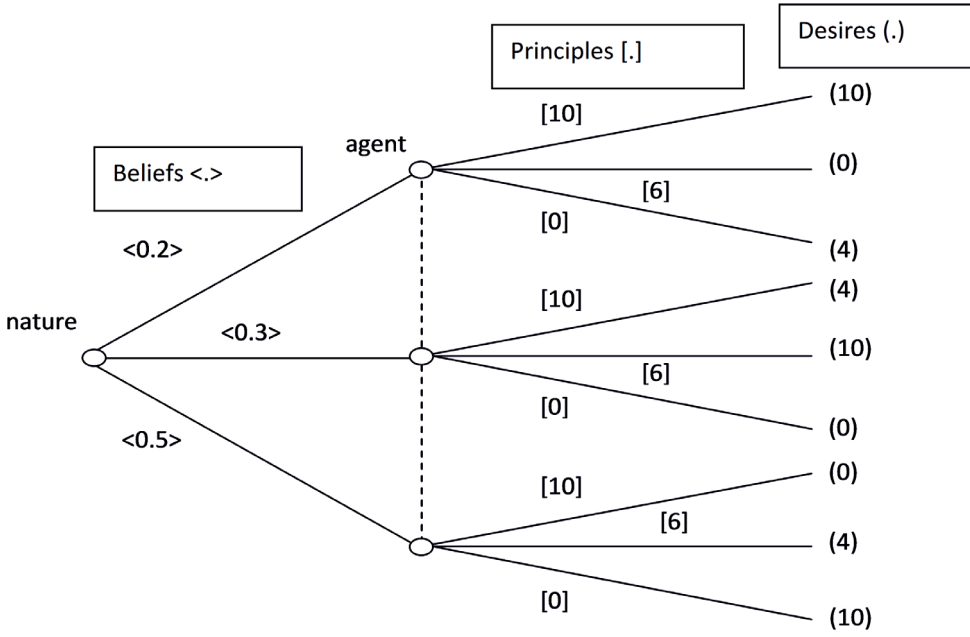
Heath sees moral norms as performing a role in decision making that is analogous to how an agent's beliefs and desires are used to prune a decision tree until only one branch remains.

> We may start out by looking at the options that are actually available to us. This amounts to a doxastic pruning of the decision tree. We then consider which among these options are normatively permissible, that is, we eliminate all the options that violate the prevailing set of social norms. This amounts to a deontic pruning of the decision tree. Once this is finished, we can select the action that gives us the most desired outcome (this is the desiderative pruning of the tree) (Heath 2008, p.74).

Heath explains the analogous relationship between doxastic modalities and deontic modalities: the doxastic modalities are possible (*M*), necessary (*N*), and impossible (*I*), whereas the deontic modalities are permitted (*P*), obligatory (*O*), and forbidden (*F*). Heath suggests that an agent first consider the possible (*M*) ways of obtaining a desired outcome. Then, out of *M*, consider the permissible (*P*) ways of obtaining the same outcome. The final decision will be the most desirable option out of the remaining set (if it is not empty).

In other words, just as states can be known with varying degrees of certainty, actions can be more or less 'morally unacceptable.' Thus, Heath suggests, 'permissible actions can be ranked as more or less probable. A forbidden action, according to this view, could be represented as having an appropriateness of say, zero, and an obligatory action as having some high positive value' (Heath 2008, p.74). This allows the agent's utility function to be enjoined with normative constraints to provide a general decision framework: 'beliefs associated with states, principles associated with actions, and desires associated with outcomes' (Heath 2008, p.75).

Figure 3



A decision tree can be used to represent this approach. Figure 3 reproduces the decision tree drawn in Heath (2008, p.75). Corresponding to the decision tree, he proposes that the expected utility function be redefined as a 'value function' over actions. Let actions be indicated by a, which are chosen at the agent's nodes (in Figure 3, there are three actions, but the three nodes are preceded by three possible states of nature). The probability of each state is represented by p (in Figure 3, they are 20%, 30%, and 50%). The utility of outcomes is represented by $u(o)$ (in Figure 3, the utilities of the same action are different under different states of nature). The normative appropriateness of each action— or 'moral principles'— are represented by $n(a)$, i.e., in Figure 3, $n(a_1)$ is always **10**, $n(a_2)$ is always **6**, and $n(a_3)$ is always **0**. The value function is thus,

$$v(a) = n(a) + \sum p(o|a)u(o) \qquad (2)$$

The value functions for the three actions depicted in the decision tree would thus be:

$$v(a_1) = 10 + [(0.2)10 + (0.3)4 + (0.5)0] = 13.2$$
$$v(a_2) = 6 + [(0.2)0 + (0.3)10 + (0.5)4] = 11$$
$$v(a_3) = 0 + [(0.2)4 + (0.3)0 + (0.5)10] = 5.8$$

Accordingly, the agent would select action a_1 [11].

Heath summarizes his approach: 'Deontic constraints are accommodated simply by integrating direct preferences over the set of actions with the expected utility derived from their anticipated outcomes' (Heath 2008, p.77). It should be noted that Heath does not settle here [12]. Indeed, the discussion following Equation 2 in Heath is a philosophically rich discussion about the possible limits of utility functions representing deontic moral norms. That said, a drawback of his approach to deontic norms presented here is the inability to 'break' such norms or express guilt or moral dissatisfaction with one's decision— as we saw in Kuran (1998).

Preference rules— Rabin (1995)

The work of Mathew Rabin is often cited for its formulation of exogenous moral norms as preferences in both a consequentialist and a rule-based (nonconsequentialist) fashion. The work is primarily concerned with the estimation processes of agents under two different moral norms, and it incorporates rule-following behavior by way of the probability of outcomes of actions. An activity (x) is not represented with a real-numbered variable, but is either done ($x = 1$) or not done ($x = 0$). It renders a level of satisfaction for an agent, $V(x) = V(1) = v \in (0, 1)$. But this activity may have a social harm, $W(x)$, which we will assume does not exist if the activity is not done, $W(0) = 0$ [13]. The activity either does cause harm ($W(1) = 1$) or does not ($W(1) = 0$). However, from the agent's perspective, there is uncertainty with regard to the harm's existence. That is, the existence of harm, $W(1) = 1$, is given a probability of q and, the existence of 'no harm,' $W(1) = 0$, is given a probability of $(1 - q)$ by the agent.

One type of agent will maximize an expected utility function that incorporates both her personal benefit from the activity, $V(x) = v$, and the expected harm to society, $qW(x)$, which enters as a negative element in the expected utility function, $U(V(x), qW(x))$. Rabin calls this type of agent a 'Preference-agent' (or P-agent); her expected utility function is

$$U_P(V(x), qW(x)) = U_P(v, q) = \begin{cases} v - q & \text{if } x = 1 \\ 0 & \text{if } x = 0 \end{cases} \qquad (3)$$

The P-agent engages in an activity if and only if $v \geq q$; that is, when the personal benefit, $v \in (0, 1)$, is greater or equal to her estimation of the probability of social harm, q. (An assumption of the model is that when $v = q$, the agent chooses $x = 1$.)

Another type of agent will have a different expected utility function: They will decide against an action if the probability of harm (q) is above some 'unacceptable' level of probability, y. This unacceptable level of probability of harm may be viewed as a constraint.

Rabin then defines a function $g(.)$ on the estimated probability of harm, $g(q)$, such that if $q > y$, then $g(q) = 1$, and if $q \leq y$, then $g(q) = 0$. This function, $g(.)$, may be thought of as a circuit-breaker on the agent's consideration of the activity. Rabin calls this agent a 'Rule-agent' (or R-agent) and their expected utility function can be stated as

$$U_R(V(x), g(q)W(x)) = U_R(v, q) = \begin{cases} v - g(q) & \text{if } x = 1 \\ 0 & \text{if } x = 0 \end{cases} \qquad (4)$$

With R-agents, if the probability of social harm is high enough (i.e. over y), then the agent does not engage in the activity [14]. Notice that this agent maximizes her personal benefit alone if the probability of social harm is estimated to be under y. With this agent, then, the willingness to satisfy one's own interest is constrained: there is no consideration of the action if the probability of harm exceeds some maximally-acceptable level of probability of harm. This is in contrast with the P-agent: they will seek their interests so long as the probability of harm does not exceed the personal benefit of the activity. But also, the R-agent benefits fully from their self-serving activity if $q < y$, whereas the P-agent only benefits from their self-serving activity by $v - q$ (assuming the action is done or $v > q$).

Despite the different decision formulations, nothing assures different behaviors. Both agents will decide to do the activity ($x = 1$) when $q < v = y$; likewise, both agents will not do the activity if $q > v = y$ [15]. In essence, under this condition, the agents are deciding according to the same decision rule: do the activity when $v \geq q$. Formally,

$$\underset{x}{\text{argmax}}\ [U_P(v, q)] = \underset{x}{\text{argmax}}\ [U_R(v, q)] \qquad (5)$$

holds for all q and v when $y = v$, or $U_P(v, q) = U_R(y, q)$ [16]. Rabin calls this equivalence of U_R and U_P under all q and v a *fixed-belief behavioral equivalence* (FBBE). It gives us the ability to predict the behavior of agents by simply knowing their beliefs (their q), without knowing their morality.

The focus of Rabin then becomes how these moralities will act when they can choose how to update their beliefs about the harms of their actions. The existence of a bias in the selection of information will lead one moral type to act differently than another moral type; specifically, Rabin demonstrates that 'when people can manipulate their beliefs [...], the likelihood that an agent will cause social harm is higher if she abides by moral constraints than if she were to abide by [...] moral preferences' (Rabin 1995, p.4). For our purposes, Rabin's work serves to highlight how preferences can represent deontic-type moral norms through one element of an action (here, the probability of harm), while remaining consequentialist in orientation through ultimately evaluating the outcomes of actions for their moral import.

Deontic Thresholds— Zamir and Medina (2010)

Zamir and Medina have as their primary concern the incorporation of deontological constraints into the standard cost-benefit analysis used in policy decisions. However, their analysis 'applies to both acts and rules, to both moral and legal questions, and to both private and public choices' (Zamir and Medina 2010, p.80). Thus, the orientation of their model of decision making is not so much the individual agent, but the public administrator or agent of social policy [17]. Their framework is nevertheless significant for our purposes, and while it may be the simplest and most direct incorporation of deontic moral norms into the economic theory of choice, it is the most extensive consideration of types of deontological morality out of the works reviewed here.

The cost-benefit analysis (CBA) Zamir and Medina espouse incorporates what they refer to as 'threshold constraints' that reflect a 'moderate form of deontology.' This is in contrast to an 'absolutist deontology,' which 'maintains that constraints must not be violated for any amount of good consequences' (Zamir and Medina 2010, p.46). Thus, a moderate deontology views constraints that 'may be overridden for the sake of furthering good outcomes or avoiding bad ones if enough good or bad is at stake' (Zamir and Medina 2010, p.46). The analysis is thus applied to a single action, where the costs (C) and benefits (B) are calculated for each action under consideration by an agent. The threshold function is represented by T, which takes a positive value when the act is permissible. Thus, the threshold function is a function of three variables: the harms or costs of the action, C, the benefits of the action, B, and the action's deontic constraint, K. The simplest example of a threshold function is what they call an *additive* function, $T = T(B, C, K) = B - (C + K)$.

A *multiplicative* function ($T = B - K \cdot C$) could be combined with the additive function, like so: $T = B - C \cdot K_2 - K_1$, which allows for simple consequentialism (when $K_1 = 0$ and $K_2 = 1$) or absolute deontology (with very high levels of K_1 or K_2), or moderate deontology (when K_1 and K_2 are at intermediate levels). Zamir and Medina suggest that the form of the constraint can be decided by its appropriateness for the issue analyzed (Zamir and Medina 2010, pp.79-104). (For example, if the constraint is to be measured against benefits and costs on a different scale, one may choose $T = \ln(B - C) - K$.)

It should be noted that only the permissibility of an act is determined by the threshold function, whereas the optimal (or morally best) amount of the act should be determined by a traditional CBA calculation. Although Zamir and Medina note that implementation of the threshold implies a two-step process [181], they do not necessarily prescribe a maximization of benefits (utility) after morally objectionable actions have been discarded by the constraint, K. For example, they suggest that in some cases, after it has been determined that an act renders 'a sufficiently large net benefit to override the constraint,' the lexical priority of one action over another may be considered (Zamir and Medina 2010, pp.83, 148). That is, while the net-benefit of one permissible act may exceed that of another permissible act, the latter may be chosen because, for example, the former is 'categorically inferior' in some morally relevant sense.

In addition, the threshold function 'not only determines the size of net-benefits required to justify overriding the constraint but also the types of benefits and costs that should be considered in this regard' (Zamir and Medina 2010, p.99). To illustrate this point, they provide a net-benefit function only for the relation between B and C, wherein each benefit is reduced (if at all) by its related cost: $B_i - C_i = b_i$. They then incorporate a threshold for each type of net-benefit, K_i, like so:

$$T = B - C - K = b(b_1, b_2, ..., b_n) - k(k_1, k_2, ..., k_n) \qquad (6)$$

We see, then, that while the relationship between the $b(.)$ and $k(.)$ functions may be straightforward (however complex the functions themselves are), the important relationship between the net-benefits and their corresponding constraints can be stipulated in a multitude of ways. Furthermore, a threshold function may apply different weights to different types of costs and benefits so as to give more importance to some benefits/costs/constraints. The range of threshold functions is therefore very broad, and their proposed method for how a deontic belief may be

incorporated into a cost-benefit framework is fairly straightforward. It is worth noting though that this model does not alter the traditional utility maximization framework in any significant way, since the CBA method is effectively retained. Avoiding this calculation approach to morality is exactly the motivation of the next author considered.

Principled restraint— Rose (2011)

Daniel Rose considers the idea of agency under the influence of moral norms in the context of a grander issue: what must the moral norms of people be in a functioning and prosperous large market economy? His starting point is a simple cost-benefit equation, which he labels the moral principle of a 'rational opportunist' (Rose 2011, pp.24-25). He defines $z(x)$ the net-payoff from taking an action, x, but the ultimate 'value' of doing the action, $V(x)$, is the result of the difference between the utility received from consuming the net-benefit, $U(z(x))$, and other costs (not incorporated into $z(x)$). These other costs might be the 'expected cost of retaliation, the expected cost of a ruined reputation, the expected cost of feeling embarrassed, the expected cost of being shamed, and the expected cost of feeling guilty' (Rose 2011, p.25). Thus, the rational opportunist's objective function is

$$V(x) = U(z(x)) - C(x) \qquad (7)$$

Rose introduces the 'golden opportunity' situation, where 'there is no chance of being detected, so costs associated with institutional sanctions, the cost of a ruined reputation, the psychic cost of experiencing feelings of shame, etc., are all irrelevant and therefore can be ignored' (Rose 2011, p.92). In such a situation, while $C(x) = 0$, morality may at least be induced by empathy for others. Thus, he invites a second agent into the model in order to demonstrate the effects of 'harm-based moral restraint' (Rose 2011, pp.92-95). Thus, we have two agents, A and B, where B is the agent considering an act that promotes their welfare, but is deleterious to A.

Rose represents B's concern about the welfare of A with two elements: B's expectation of harm done to A, $E[\Delta U_A(x)]$, and some measure of the intensity for which B sympathizes with A, θ_A. In addition, Rose separates the intensity of concern for A's welfare from the amount of guilt that B suffers if the action is undertaken despite the care for A. The guilt-function, $g_B(x)$, is thus multiplicative of the concern B has for A's welfare, where $g_B(x) \in [1, \infty]$. The resulting expression from B's perspective is thus the harm-based value function:

$$V_B(x) = U_B(z(x)) - g_B(x)\theta_A E[\Delta U_A(x)] \qquad (8)$$

The answer to the question of whether B will act against the interest of A reduces to whether or not the first term exceeds the second. Rose points out that there is no difficulty in incorporating concern for others into the model of rational opportunism. There is no effective difference between the agent not choosing the action because of $C(x)$ in Equation 7, and not doing the action out of concern others in Equation 8. This is concern for Rose because of his grander argument that moral restraint based on harm or empathy is unlikely to prevent opportunistic behavior in large groups, since people's connection with others is reduced in such contexts (Rose 2011, pp.98-102).

An alternative type of moral restraint would be what Rose calls 'principled moral restraint,' where agents choose to act in accordance with what they believe to be morally right regardless of the harms (or ancillary social costs) produced by the act (Rose 2011, p.108). In such a model, according to Rose, the agent is not rational if they undertake an action for some benefit, when they believe it to be wrong (and therefore experience a greater displeasure from guilt). This guilt is a different type of guilt than that introduced above (for Equation 8), which was attached to the expected harm done to another. This second type of guilt may be thought of as 'principled' guilt.

Rose therefore offers a new value function to capture the two different types of guilt:

$$V(x, n) = U(z(x)) - g(x)h(x, n) - g_I(x) \qquad (9)$$

where $g_I(x)$ is guilt attached to the act itself. Rose proposes that, since the harm of the action is a function of the number of people affected, and harm is diluted by increasing the number of people, Equation 9 can also reduce to something like Equation 7 (Rose 2011, p.108-111). Formally, $g(x)h(x, n) \to 0$ as $n \to \infty$, such that, for large populations, we have $V(x, n) = U(z(x)) - g_I(x)$. In other words, in large groups, principled moral restraint is the only guard against opportunistic behavior, and in cases where there is a large net-payoff, $U(z(x))$, it will not be a sufficient moral restraint[19]. Nevertheless, we see that $g_I(z)$ has a very similar effect to that of K in the threshold function of Zamir and Medina (2010), and acts in a similar fashion as the deontic constraint of Kuran (1998).

Discontinuous goods— Dowell et al. (1998)

The last work worth mentioning is that of Dowell et al. (1998). Their presentation of the influence of moral norms on decision making aims to capture a dual-nature

of moral norms: their ability to be traded for other goods in some cases, while in other cases they resist this. At times, the existence of moral norms would seem to infer preferences that are discontinuous or 'lumpy,' while at other times agents seem to exchange a moral action for non-moral activities. The authors start by adding a 'moral value' for honesty (H) to the typical utility function:

$$U = U(X_1, X_2, \ldots X_n, H) \qquad (10)$$

They stipulate that H is not continuous: it takes the value of 1 when the agent is moral, and 0 when the agent is immoral, or chooses to be dishonest. This captures the case where there is no continuous trade-off between H and the other elements of the utility function, X_i.

> Rather than treating moral decisions as being marginal trade-offs, analogous to choosing to tell one more lie in return for three additional apples, we treat them as analogous to the 'lumpy' nontrivial moral decisions of choosing whether to become a drug dealer.

This also implies that being moral increases the agent's utility from any given basket of goods.

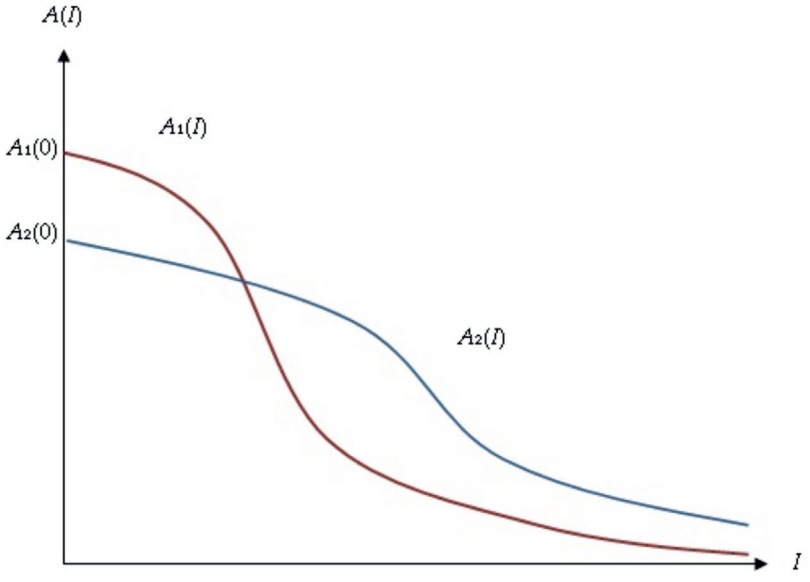
Their extended model attempts to allow for both a continuous and discontinuous standing with regard to the moral value of actions.

They posit a function, $A(.)$, over the 'immorality' of an action, I, whatever the action may be, and include it in the utility function like so:

$$U = A(I) \cdot f(X_1, X_2, \ldots X_n) \qquad (11)$$

In this case, $A(.)$ decreases in I, so that increasing the level of the immoral activity diminishes one's overall utility. The authors represent $A(I)$ as a cubic function resembling that in Figure 4, which is reproduced from Dowell et al. (1998).

Figure 4



The positive affect of being moral is effectively the same as in Equation 11: the agent's utility is increased by being moral given any vector of goods, *X*. However, in Equation 11, *I* is not alongside other goods, *X*, nor a function of the other goods. The importance of $A(I)$ is that it can represent a rigid or discontinuous morality (like *H* in Equation 10), or it can represent a less-rigid and more continuous morality that, for example, does not reduce one's utility drastically with 'a little immorality' (as might resemble $A_2(I)$ in Figure 4). In other words, a 'less-critical' immoral action would extend $A(I)$ further over *I*, while a 'severely immoral' act can be represented by both an $A(I)$ starting much lower (on the vertical axis of Figure 4) and rapidly decreasing over *I* (as might resemble $A_1(I)$ in Figure 4). This simple representation of a moral norm is versatile and robust in that it can accommodate a wide range of personal moralities (consequentialist or deontological), and allows one to relate them to a 'bundle of goods,' the typical elements of economic decision making.

We will now review some of the qualities of the models examined, and consider incorporating the most veracious qualities of each into a single model.

Review and amalgamation

Review of the models

We will start part by comparing and contrasting the features of each of the models. Figure 5 highlights some of the qualities of each model in terms of four key characteristics: (1) the objective function, (2) the choice variable, (3) the relation of the norm to the agent's utility, and (4) the means by which the norm influences the choice of the agent.

Figure 5

	Functions	Choice Variable	Norm's Influence on Utility	Means of Influence
Rabin(1995)	$U_P(V(x), qW(x)) = v - q$ $U_R(V(x), gW(x)) = v - g(q; y)$	x ∈ {0, 1}	Negative, if the act is harmful to others OR Negative, if the act violates the moral standard of harm to others	Utility is reduce by the amt of expected harm, $(v - q)$ OR Utility is reduce by the amt of harm, $(v - 1)$
Heath(2008)	$v(a) = n(a) + \sum p(o \mid a)u(o)$	$a \in \{a_1, a_2, \ldots a_n\}$	Positive, if the act is morally valuable OR Negative, if the act is morally unacceptable	Act given more/less weight by $n(a)$
Kuran(1998)	$U(\mathbf{x}) = I(\mathbf{x}) + M(\mathbf{x}, \mathbf{x}_m)$	**x** = (a,b) ∈ A x B	Negative, if the choice does not satisfy the moral standard	Utility reduced by some amt if $x \notin X_m$
Zamir & Medina(2010)	$T(x) = B(x) - C(x) \cdot K_1 - K_2$ Or $U = U(x, \max[0, T(x)])$	Assumed to be x ∈ {0, 1} or x ∈ $\mathbb{R}^+$	Utility is unaffected if the benefit does not exceed the costs/constraint OR Positive, if the benefit exceeds the cost/constraint	Utility is affected by the action only when T ≥ 0
Rose(2011)	$V(x, n) = U(z(x)) \ldots$ $- g(x)h(x, n) - g_P(x)$	x ∈ {0, 1} or x ∈ $\mathbb{R}^+$	Negative, if the action produces guilt in the agent from either (1) Harm done to others, or (2) Breaking one's own moral principle regarding x	The agent's sense of guilt from either $g(x)$ in conjunction with $h(x,n)$, or $g_P(x)$, or both
Dowell, Goldfarb, & Griffith(1998)	$U(X, A(I)) = A(I) \cdot f(X)$	$I \in \mathbb{R}^+$	Reduces the utility of any given basket of goods, $f(X)$	$A(.)$ falls in I at different rates, depending on the action and its level

The first column of Figure 5 gives the utility functions of each model. The various authors may have referred to their agent's utility functions as 'value functions', 'expected utility functions', or something similar, but all of them are maximized through the selection of a choice variable, and so they will be referred to as simply 'utility functions.' In the case of Rabin (1995), two different utility

functions are given, but the one of most interest is that pertaining to the rule-abiding agent, U_R, since it is the instance more reflective of deontic moral norms. Also, the utility function given for Zamir and Medina (2010) in Figure 5 is one that is implicit in their discussion of the threshold function; that is, they suggest that the agent optimize over the set of actions that pass the threshold function. The second column of the table in Figure 5 lists the choice variables of each author's function. These range from a simple binary variable (Rabin 1995) or bundle of activities (Kuran 1998).

The third column identifies how the norm influences personal utility. Usually, models represent the satisfying of (breaking of) a norm as increasing (decreasing) one's own utility; but in some cases, abidance to a norm has no impact on personal utility (Kuran 1998). The last column gives the means by which this influence is conveyed upon utility. There are a wide range of mechanisms for representing this, even more than those given in the models reviewed here. However, the models reviewed provide a good sample of the possibilities. It should be obvious, though, that the general mathematical options are limited to some formulation (and/or combination) of the operations of addition, multiplication, and exponentiation. The meanings of these mechanisms in the context of decision making under the influence of moral norms is what makes a model pertinently veracious.

Having the key elements in Figure 5, we can now inquire as to what aspects of each model are intuitively appealing, and/or if these qualities overlap with the other models.

The contribution of moral action to utility

The first point worth noting is how moral action is valued by each function. We may want to posit that the satisfaction of a moral rule brings some positive contribution to one's own utility, but there does not appear to be any a priori necessity for incorporating this into a model of moral norms. This quality is explicitly included in Dowell et al. (1998) and Heath (2008), while Kuran (1998) suggests that there be *no* satisfaction received upon the satisfying of a moral constraint. But as we will see, it is easy enough to include: there is no obvious reason why $M(x, x_M)$ must take on the value of zero if the constraint is met.

In the case of Rabin (1995), this quality cannot be easily incorporated; for example, we may stipulate that if $q < y$ (for U_R agents, by Equation 4), then g adds to v. However, the meaning of this would be curious: the agent is somehow

taking pleasure in someone's harm, even though the probability of harm is under a (personally) acceptable level of harm. This element of moral behavior, therefore, does not seem workable with Rabin's model.

In the case of Rose (2011), adding this aspect of moral behavior is plausible: we may simply add another term (to Equations 8 or 9) that is a function of *not-doing x*, so as to potentially counter the utility from the action, *x*. Alternatively, we may simply incorporate within $U(z(x))$ the lost utility of not-doing *x*. In the case of Zamir and Medina (2010), there is no positive contribution to utility from acting according to moral norms. The action has its own benefits that must weigh against the costs and exceed the constraint. The positive contribution to utility from the satisfaction of a moral constraint would therefore seem to be most easily incorporated into a reformulation of Rose's equations, or something like Heath's $n(a)$ or Kuran's $M(x, x_M)$.

Reductions in utility vs. restrictions of action

If deontic moral norms restrict action and effectively act as constraints on behavior, then it does not seem adequate to represent these norms as simply reducing the utility of an agent. In other words, if a constraint is binding, it would seem that it must very well restrict actions or reduce their possibility significantly in order to be effective. One might reply that moral constraints are not equivalent to objective constraints, since a moral rule can be broken. Thus, one might say, if it is to be incorporated into an agent's utility function, this deontic moral norm may reduce the utility of an action fully (possibly to negative levels) when not satisfied, yet remain breakable by leaving the action to be part of the opportunity set.

The choice of whether to render an action possible or impossible by way of the deontic moral norm may come down to the topic of research and what experimental evidence one has for both positions in that context. In other words, the extremity of an agent's belief about the obligatory/forbidden nature of the action may be represented by either some term that serves to reduce utility (in proportion to that belief), or by a term that serves to exclude the possibility of an action. However, it should be noted that choosing the latter option for one's model precludes the possibility that people exhibit degrees of belief about the immorality/morality of an action, an implication that might be easily invalidated through empirical investigation.

In the models reviewed here, we have examples of both intuitions. Rabin (1995) assumes that the maximum utility receivable is 1, and the violation of a norm is a deduction of 1, where an action is thereby barred from choice. Heath (2008), likewise, bars actions from choice through setting $n(a) = 0$. Kuran (1998) does not explicitly connect the utility of an option ($I(x)$) to the reduction of utility. In fact, while disutility is discussed, there is no stipulation of how the reduction of utility will take place. In the case of Rose (2011), it is clear that utility is reduced by an amount of guilt, whether it be from breaking one's principles ($g_I(x)$) or from the harm done to others ($g_B(x)$).

In the case of Zamir and Medina (2010), the deontic constraint reduces utility and, as noted earlier, can represent absolute deontology if the deontic threshold is set high enough. However, in their formulation, an action can also be prohibited, because the benefits of it are too small. In this case, it is not so much that the deontic moral norm reduces the utility of the action, but the utility of the action may not render it permissible. Thus, the prohibitive nature of the deontic moral norm is equally determined by the utility of the action in Zamir and Medina's formulation. Nevertheless, both the reductive and prohibitive quality of deontic moral norms can be represented in their model.

Dowell et al. (1998) is the most illustrative of the differences on this point, as their formulation allows for both a reduction and prohibition of actions. When $A(I)$ resembles those of Figure 4, the immorality of the action reduces the utility from a choice. On the other hand, if $A(I) = 0$ for all $I > 0$, then this action is prohibitive. Yet it is prohibitive in a peculiar manner: the immorality of an action nullifies the utility of any accompanying bundles of goods (X_1, X_2, ...X_n). This is plausible in terms of its ability to represent deontic moral norms as prohibitive (or limits of the opportunity set). However, it suggests that any level of the immoral action, while completely value-independent of the bundle, completely determines the utility of the agent. It would seem that this curiosity emerges from the mathematical representation, while nevertheless rendering what a prohibitive norm's effect on choice looks like.

That which is chosen in a moral action

The second column of Figure 5 gives the choice variables of each model. There is a variety of formulations among the models in this respect, but there are three general orientations.

First, in common discourse, we may say that an action is chosen or not-chosen. In representing this, we may designate the choice variable to take a value of **0** or **1** only. Such is the case in Rabin (1995), but we may interpret Zamir and Medina (2010), Rose (2011), or Heath (2008) as using this approach, as well. For example, Zamir and Medina do not specify the choice variable's set of membership, but they seem to suggest that the costs/benefits/constraints apply to an action without regard to its level or amount.

An alternative approach is to regard the choice variable as an element of a real-valued space, which is to say that the action has a quantity or level. For example, in Dowell et al. (1998), the choice variable is $I \in \mathbb{R}^+$. This approach is probably the most familiar to economists, and most explicit in Kuran (1998), where the action is $\mathbf{x} = (a, b) \in A \times B$, where $A, B \in \mathbb{R}^+$. This approach is accommodating to the formulation of a traditional utility function. It should also be noted that while Zamir and Medina (2010) and Rose (2011) may be interpreted according to the first approach, this second approach may be incorporated into their models without complication.

A third approach would be to regard the action as a member of a finite set, unordered and undefined. For example, in Heath (2008) the choice variable is simply $a \in \{a_1, a_2, ..., a_n\}$. This approach is the most general and adaptable to other frameworks. That is, we may restrict Heath's choice variable to $a \in \{a_1, a_2\}$, such that $a_1 = \mathbf{0}$ and $a_2 = \mathbf{1}$ as in Rabin. Additionally, we may define the elements of a so that they can be translated into the framework of Kuran: we can interpret Heath's action (a_i) as an ordered pair of variables, for example, $a_i = (x, y) \in X \times Y$.

In the next section we will generally use Kuran's approach for the choice variable [20] which, as noted above, is shared by Dowell et al. (1998). However, we will employ Dowell et al.'s approach for a different aspect of the model in the next section.

The moral relevance of uncertain outcomes

The models of Rabin (1995), Rose (2011), and Heath (2008) incorporate the uncertainty involved with actions, which is a popular extension of the model of consumer choice in economics— so much so that it is now part of the canon of the theory of choice in economics. In incorporating uncertainty, each author articulates what amounts to an 'adjusted' expected-utility function [21].

In the case of Heath, the estimated probabilities of outcomes are according to states of nature that affect the utility of the outcome for the acting agent. In the case of Rabin, the probabilities of outcomes pertain to some harm being done to others. And in the case of Rose, the expected outcome is also an expected harm done to another party. Thus, in the models of Heath and Rose, the probability is not directly an object of moral importance, in contrast to Rabin. In the former models, the probability (whether subjective or not) functions much like it does in the typical expected utility function. In Rabin's model, since the level of harm is equal to the maximal level of utility, the resulting utility is an expected utility, but the normative standard is explicitly a particular level of probability, and not a function of the action as in Heath or Rose.

The uncertainty inherent in the outcomes of actions, as well as one's estimation of the probable harm of actions, is an aspect of actions worthy of moral consideration. Many normative standards are based on the probability of uncertain outcomes of actions, such as the legal concepts of reckless endangerment and gross negligence. A lack of foresight or disregard for common estimates of harm can impute culpability. Furthermore, even in the case of models like Heath and Rose, the weighting of utility (or moral value) by probability estimates adds realism to a model that aims to represent the influence of moral norms on choices. Moral deliberation is often concerned with the likelihood of harms from actions. It therefore makes sense to incorporate these two elements— harm to others and one's estimate of the probability of an outcome— into a model of moral norms.

A comprehensive formulation

The qualities of the models noted in the previous section suggest that several aspects of moral norms should be present in a comprehensive formulation of a model of choice: (1) the personal satisfaction experienced from satisfying one's moral constraints and/or contributing to the welfare of others via fulfillment of one's obligation; conversely, the reduction in one's personal satisfaction due to either the failure to meet their own moral standards or the negative impact of their actions on others; (2) the decision against an action as a type of action— or the categorical rejection of actions— whether due to some quality of their outcomes or not; and (3) the ability to ascribe moral value to choices with uncertain outcomes. These general concepts will guide the formulation of a comprehensive model.

The choice variable

An action has a multitude of ways of being characterized. Even the naming of an action, or identifying the nature of the action, can be perplexing. For example, the act of exchange may characterized as purchasing a product by the buyer, but it may also be regarded as consuming the product, or simply 'choosing' a product in the simple consumer choice model. However, under moral scrutiny, these might be completely different actions. Furthermore, what matters about the action of purchasing a product may be the amount consumed, the consequences following, the price paid/charged for it, the use of the product, or even the intention of the buyer. Thus, any action can have several elements, any number of which— independently or collectively— may have moral importance. All of these factors likely depend on the issue and scenario being studied, but a model of general applicability should allow for alternative formulations.

Therefore, to begin a basic formulation of such a model, I will regard the action as a choice of an ordered set in $\mathbb{R}^N$, $\mathbf{a} = (a_1, \ldots a_n)$, which has the benefit of allowing an action to have a multitude (n) of morally relevant characteristics. Thus, a comparison of two actions $\mathbf{a}_1$ and $\mathbf{a}_2$ is a comparison of vectors in $\mathbb{R}^N$. This is similar to Heath's $\mathbf{a} \in \{a_1, \ldots a_n\}$ except that now each $a_i \in \mathbb{R}$. Thus an action is a set of elements, each of which can vary in magnitude.

For the sake of simplicity we will only refer to an action as an ordered pair, $a = (x, y) \in X \times Y$, allowing an action to have two elements of moral worth. Both x and y of the single act, a, can be regarded as having a moral aspect in themselves, yet they can also have a moral aspect as a part of a combination. Also, in some cases it may not be realistic that each element of an action be represented by a real-valued variable, and in some cases it may make sense to regard each $a_i \in \{0, 1\}$. It is therefore logically possible that one quality of an action be $x \in \{0, 1\}$ and $y \in \mathbb{R}^+$. Another general characterization of actions would be to call an action $\mathbf{a} = \{a_1, \ldots a_n\}$. Indeed, some authors (Suzumura and Xu 2009) only employ the standard utility function, making use of the set of real numbers, in relating consequentialist and non-consequentialist considerations. However, since we have considered models that express deontic moral norms explicitly in the form of utility functions, the usual assumption about the potential option set will be maintained and $X, Y \in \mathbb{R}$.

The moral opportunity set and inherent moral value of actions

Borrowing Kuran's notion of a moral opportunity set, X_M, it seems reasonable to define a set or range of morally permissible action, determined by both lower-bounds ($\underline{x}$, $\underline{y}$) and upper-bounds ($\bar{x}$, $\bar{y}$). Thus, a moral opportunity set will have the structure

$$A_M = \{(x, y) \mid \dot{X} \geq x \geq \underline{x}, \bar{y} \geq y \geq \underline{y}\} \qquad (12)$$

The basic function of the moral opportunity set is to establish the constraint-like nature of deontic morals. It effectively defines an action (or qualities or actions) as morally relevant. Of course, A_M by itself says nothing of the impact of choosing $a \notin A_M$. But it establishes the possibility of moral norms instructing 'absolute prohibitions' against actions ($\dot{x} = \underline{x} = 0$ and $\bar{y} = \underline{y} = 0$) or against the choice of one of its elements ($\dot{x} = \underline{x} = 0$ and $\bar{y} \geq \underline{y} > 0$). The structure of the moral opportunity set is obviously a construct of the binary relations of equality and inequality, which function well on real-valued variables. If we were to use a different space for the definition of actions, a moral opportunity set may not be as easily defined.

To incorporate the inherent moral value of actions, as represented with $n(a)$ in Heath's utility function, a function resembling Kuran's moral value of an act, $M(x, x_m)$, will be used. As noted above, it seems reasonable to let this function take both positive and negative magnitudes, as with Heath's $n(a)$. In essence, the moral value function will represent the value an agent puts on an action without regard to its outcomes. Additionally, in Kuran's formulation, the moral utility of an action is a function of the action itself and the moral opportunity set, but this aspect of Kuran's will not be followed. Let the inherent moral value of the action for the agent be represented by $M(a, A_M)$ or $M((x, y), A_M)$.

A questionable implication of including the moral opportunity set in the representation of the inherent moral value of an action is that the moral value of the action and the utility received from satisfying this moral norm are one and the same. Conceptually, a difference may exist between the two notions, but this may not be of importance for the explanation of choice. It does not seem inconsistent to say that the utility (disutility) derived from satisfying (failing) one's moral standard is equal to the value one puts on performing (not performing) the action in a given context.

Rose also combines the moral opportunity set and the inherent moral value of an action, since the action considered is assumed to be 'morally negative', and harmful

to another party while beneficial to oneself. The harm done to the other party also reduces one's utility through their guilt function, $g(x)$. Thus, choosing an action outside one's moral opportunity set renders a disutility in proportion to how intensely one feels guilt about not abiding by their morality. The function $A(.)$ in Dowell et al. could be interpreted as the complement of the moral opportunity set $M(a, A_M)$. Nevertheless, $A(.)$ gives both a disutility from choosing from a set of actions (I), and a reduction in utility derived therefrom. Thus, in Dowell et al., the inherent moral value of an action is tied to the disutility of not refraining from it.

Other-regarding preferences and uncertain harm

As seen in Figure 5, the personal utility of an agent due to their action is represented in Heath's model by $u(o)$, in Kuran, by $I(x)$, in Rabin, by $V(x)$, in Rose, by $U(z(x))$, in Dowell et al., by $U(A(I), X)$, but only in Rabin and Rose is the utility of others explicitly represented. The importance of moral norms often shows itself in cases of strategic interaction or the mutual dependence of outcomes and welfare between agents, and a popular notion of moral norms is altruism or regard for others' welfare, commonly captured by other-regarding preferences (see Bowles 2003, Ben-Ner and Putterman 1998, Falk and Fischbacher 2005). We therefore include, separate from one's 'pure self-interested' utility, the welfare of others from an outcome of an individual's choice of action. In the case of both Rabin and Rose, the outcome of an action was harm to the other, which reduces the utility of an agent according to some function. However, there is no reason to not include the possibility that one's utility increases with the welfare of another. This single addition will be done by letting the measure of others' welfare take both positive and negative values, such that the latter reduces one's utility.

The second issue regarding outcomes is their uncertainty. As we saw, all of the models reviewed (with the exception of Kuran and Dowell et al.) incorporate this element. It is most explicitly drawn out in Rabin, but Zamir and Medina also analyze this idea in parts of their work, where the probability of the outcome is an element of moral import. That is, the probability of harm to another is itself, aside from the amount of harm, the factor determining the moral worth of the action. This is an important consideration as moral rules often as expressed in (legal) terms like reckless endangerment, gross negligence, or foreseeable risk. It therefore makes sense to let the morality of an action be determined not only by the welfare of an outcome, but also by the probability of that outcome.

The outcomes of actions are weighted by their probabilities in the models of Heath, Rabin, and Rose. In a similar fashion, I will regard the outcome (o) of an action as having a probability, $P(o|a)$, but this will weigh on both the agent's utility, and the welfare of others from this outcome (thereby incorporating other-regarding preferences). No assumption will be made with regard to the distribution of $P(.)$ over o, nor its subjective/ objective status; but we obviously need to assume that $P(.)$ is additive over o for each a, thus requiring each o to be independent.

Let the agent's personal utility be represented by $S(o)$, and the welfare of others due to this outcome by $W(o)$.There is no reason why $W(o)$ cannot represent a change in welfare of others (as opposed to the static level of welfare of others), as in Rose. In some cases the change in welfare may be more relevant than the resulting state of welfare. Furthermore, the 'others' here is 'all parties affected by outcome o,' and $W(.)$ may be constructed for unequal weights among them. One's expected utility from an action [22], to incorporate the possible harm to others, will be represented as the sum of the products of the probability of the outcome itself on the difference between $S(o)$ and $W(o)$, as such:

$$\sum P(o_i \mid a) \cdot [S(o_i) + W(o_i)] \qquad (13)$$

To represent various degrees of concern for others (e.g. altruism versus self-interestedness of agents) we may make the difference in $S(o)$ and $W(o)$ a weighted average, like so

$$\sum P(o_i \mid a) \cdot [(1 - \alpha) S(o_i) + \alpha W(o_i)] \qquad (14)$$

where α would represent a degree of altruism or importance of others' welfare in one's own satisfaction from actions, comparable to Rose's term, $g_B(x)\theta_A E[\Delta U_A(x)]$.

To include the idea that the level of the probability of an outcome is of moral importance by itself (as in Rabin), we might suggest that the probability function be expressed as a function of not only the outcome (o) and action (a) but also some threshold of likelihood (q), like so $P(o_i|(a, q))$. However, this would be inadequate: the variable q should presumably be of such importance that it determines the value of the entire expected utility. For example, we want to stipulate that if the probability $P(.)$ of a harmful outcome ($W(o) < 0$) is greater than q, then the action should not be done— that is, the fact that $P(.) > q$ for one outcome (o_i) renders negative expected utility for the action (a) over all outcomes. Thus, the relationship of q to $P(.)$ must be expressed as a function of the welfare of others ($W(o)$), and independent of the calculations presented in Equation 13 or 14. We can let

$q = q(W(o))$ be the element of an identity function $I(.)$ that takes the value of **0** when $P(.) > q$ and include it in Equation **13** like so:

$$I(P(o) > q)[\sum P(o_i \mid a)\cdot[S(o_i) + W(o_i)]] \qquad (15)$$

Let me summarize what has been proposed with a more general consistency in notation: each action $a \in \mathbf{a}$ produces outcomes $o_j = \{o_1, \ldots o_m\}$, and each outcome o_j has a probability $P(o_j|a)$ that is related to a threshold q, which is a function of the welfare of others under each outcome, $W(o_j)$. We might reasonably instead make q a function of total welfare of others over all outcomes, so as to suggest that what matters is not the welfare of others with regard to one specific outcome, but the welfare of others over a range of outcomes from the action; that is, make q a function of $W(o)$ and $P(o)$ instead of $W(o_j)$ and $P(o_j)$. It might observed that a relationship between q and the function $M(a, A_M)$ outlined in the previous section can be reasonably introduced, since they both render discontinuous, constraint-like, or absolutist qualities for the decision of the agent. In other words, q defines a moral opportunity set in the space of uncertain outcomes. They might very well be related in the mind of the agent, but we will keep them as separate terms going forward.

Properties of the comprehensive model

In sum, we are suggesting an expected utility of an action with the conceptual form like the following (using Equation **13** instead of **14**):

$$U(a) = U(x, y) \qquad (16)$$

$$= M((x, y), A_M) + I(p, q)[\sum p(o \mid (x, y))\cdot[S(o) - W(o)]]$$

This equation reduces to the standard expected utility function of homo-economicus (or Rose's opportunist value function) when $M(a, A_M)$ is an empty set, when $W(o) = \mathbf{0}$, and $I(.) = \mathbf{1}$. We can conversely say that a morality is captured by these three elements, and the formal specification of these elements gives the full character of the moral norm.

Consider the term $M(a, A_M)$. There are a multitude of possible formulations for this term, but the expression should depict the sense in which a moral norm values a set of actions (or quality of actions) by themselves, without regard to the outcome. Wishing the term $M((x, y), A_M)$ to take on both positive and negative values, an obvious suggestion would be for it to be positive when the action satisfies a moral

opportunity set constraint, $(x, y) \in A_M$, and non-positive when this is false. Also, it should be clear that a categorical prohibition against an action— for example, x— would be equivalent to where the upper-bound on x is equal to zero.

Example 1 (*Uniform Valuation of* A_M)

A morality may require that only $a \in A_M$ are valuable. This implies that the complement of the moral opportunity set, A^C_M, is devalued as a class. We may say that two extremes of this are an absolute devaluation, where $a \notin A_M$ renders $M(a) = 0$, or relative devaluation, where $M(a', A_M) < M(a, A_M)$ for $a' \notin A_M$.

Example 2 (*Devaluation of the Complement of* A_M)

Another possibility is to treat the A_M as not particularly valued, but the complement devalued. In fact, we may use something analogous to the immorality function, $I(.)$ in Equation 11 from Dowell et al. (1998), to describe the value of $a \notin A_M$. That is, the 'further away' one is from A_M, the less value the action has. Assume for simplicity that the action is a single-variable, $a = x \in \mathbb{R}$ and that $x = \dot{x}$ has maximal moral worth of z; all other levels below $\dot{x}$ are valueless, and all levels above are devalued according to a function resembling $A(I)$ of Figure 4. In this case, the moral opportunity set is a singleton, $A_M = \{\dot{x} \mid \underline{x} = \dot{x} = \ddot{x}\}$ and $M(x, A_M) = z$. Then our term $M(a, A_M)$ is

$$M(x, A_M) = (\max[0, (x - \dot{x})/(|x - \dot{x}|)]) \cdot f(x) \quad (17)$$

where $f(x) = \lambda_3(x - \dot{x})^3 + \lambda_2(x - \dot{x})^2 + \lambda_1(x - \dot{x}) + z$ (and the parameters approximate the values of $\lambda_3 \approx 1.35$, $\lambda_2 \approx -2.25$, $\lambda_1 \approx 0.20$, and $z \approx 0.8$). Thus, $x \notin A_M$ are less valued (if at all), and are devalued more the 'more immoral' they are.

Example 3 (*Virtuous Mean*)

Another possible specification of $M(a, A_M)$ would set the disutility of $a \notin A_M$ equal to the 'furthest' transgression, while the positive contribution to utility from meeting one's moral opportunity set is only equal to the closest boundary (or 'lowest moral achievement'):

$$M((x, y), A_M) = \min[((\ddot{x} - \underline{x}) - \max[|\underline{x} - x|, |\ddot{x} - x|]), \quad (18)$$
$$((\bar{y} - \underline{y}) - \max[|\underline{y} - y|, |\bar{y} - y|])]$$

This expression implies that the highest utility of the moral opportunity set is achieved at the midpoint between the upper and lower moral bounds of A_M. One counterintuitive consequence of this formulation is that the positive contribution

to utility is smaller for 'smaller targets'— i.e., smaller A_M It may seem that the converse should be true. Likewise, one's overall utility $U(a)$ would be reduced by selecting from outside the moral opportunity set.

The second and third terms, $W(.)$ and $I(.)$, are inexorably linked. First, it is clear that $I(.)$ is relevant so long as q is an effective boundary on $P(.)$. For example, if $q = 1$, then $I(.) = 1$ and becomes a redundant term. Secondly, as mentioned above, q is a constraint with regard to possible effects of actions, and may be related to A_M However, keeping them separate has an intuitive appeal. When $q < P(o)$ and $I(.) = 0$, the action may still be chosen, since— while the second term of Equation 16 is 0— the value of $M(.)$ may be positive.

The implication of this is that the action is not necessarily decided against when $P(o) > q$. This, however, is not necessarily a counter-intuitive outcome: It may capture the 'conflict of obligations' that people do not uncommonly experience in choosing according to moral norms. In fact, as Equation 16 suggests, these inner conflicts are often between meeting an obligation in the face of disagreeable effects from actions. It may be fitting then to keep the probability limit, q, relevant to only the second term of Equation 16.

Concluding remarks

Equation 16 may be useful as a first approximation of an agent's moral disposition towards an action, incorporating several sensibilities behind moral behavior: the sense of moralities as rules or constraints, moralities as personal sensibilities and sensitivities towards conduct and outcomes, and moralities as concerning the impact of one's actions on the welfare of others. It is also general enough in its formulation to be applicable and illustrative of agents' possible behavior under the influence of a given set of moral beliefs.

The above discussion of the attempts to model moral norms has focused on moral norms not as formal structures of conduct between members of a group, but as personal factors held by individuals that influence their actions. The prospective model was one that could be applicable to individuals in general, regardless of their individual morality, yet provide a formal framework for representing any individual morality. The motivation behind the review and amalgamation of models was thus both to help explain human action, and portray a convincing set of conceptual relationships that are true to our intuitions about the influence of morality, especially in the form of deontological moral norms.

The outcome we reached is a formal representation of decision making under the influence of moral norms that achieves a greater degree of generality than any of the individual models reviewed. The robustness of the *tout-ensemble* model (Equation 16) lends itself for further development and research. Further development of this line of inquiry may be the testing of the model specifications in real-world scenarios. For example, we might test the degree to which ostensive moral behavior is rule-based or consequence-based, and whether it is other-regarding or not. The results of such an inquiry would naturally lead to a finer development of the model and possibly a new specification of this general formulation.

Acknowledgments

I am grateful for the comments of Duncan Foley, William Milberg, and Lopamudra Banerjee on an earlier draft of this paper.

Endnotes

[1] Some recent examples of such endeavors can be found in Bicchieri (2005), Fehr and Schmidt (1999), Falk and Fischbacher (2005), and Bowles and Polania-Reyes (2012). See Bowles (2003) for more examples of this line of research. Furthermore, in this paper, we will use the term 'model' loosely, as it will usually only refer to the formulation of a single utility function, and not a program or complete set of equations to be solved.

[2] For example, several works of the mid-1980s— Etzioni (1986, 1988), and Hammond (1986, 1988)— were preceded by Sen (1973, 1974, 1982), and followed by others (Vanberg 1988, Anderson 1987).

[3] In this paper, moralities will simply be divided into consequentialist and nonconsequentialist, where the latter will be synonymous with 'deontological' or 'deontic.' For a summary of some deontological theories, see Zamir and Medina (2010).

[4] There is a considerable amount of literature that addresses the philosophical and methodological difficulties. Some such works are White (2011), Rose (2011), Vanberg (1988), Goldfarb and Griffith (1991a, 1991b), Koford and Miller (1991), Sen (1977, 1982, 1987, 1998), Etzioni (1988), Anderson (1987), Broome (1992), Ben-Ner et Putterman (1998), and Heath (2008).

[5] For the difficulties related to the logic and mathematics of decision theory, see Sen (1973, 1974, 1997), and Suzumura and Xu (2001, 2003, 2009).

[6] Goldfarb and Griffith (1991a, 1991b) give a different taxonomy of the approaches. Basically, they suggest that the 'choice variable' approach be thought of as a strategy in a game situation.

[7] Kuran (1998, p.232fn1) states that this is 'broader than the concept of metapreferences,' as developed by several writers, most notably Sen (1974). Also, he gives several definitions, but generally, Kuran's use of 'values' is similar to the use of 'moral norms' here.

[8] The ranking of morally feasible options is implicitly rejected by Kuran's formulation, since, if x_m is met, then the moral utility is zero. However, an intuition about morality is that there could be a comparison between ($\underline{a} \prec a$, $\underline{b} \prec b$), which might be the morally worst case, and other sets, such as ($a \prec \underline{a}$, $\underline{b} \prec b$) and ($\underline{a} \prec a$, $b \prec \underline{b}$), which may be less bad or simply better. But this is not taken up here by Kuran.

[9] It might be worth noting that Kuran rejects the dichotomy between norms as preferences, and norms as constraints. He notes Rabin's (1995) categorization of approaches, where models fall squarely into one of the two types.

[10] Heath also quotes Peter Hammond (1998, p.25): 'an almost unquestioned hypothesis of modern normative decision theory is that acts are valued by their consequences.' Thus, a rule cannot 'confer value on an action,' but, according to Heath, 'they have tried to show that concern for rules is just a subset of concern for consequences, misleadingly described. The major motivation for these efforts has been a rather diffuse sense that expanding the conception of practical rationality to include non-instrumental reasons for action would involve introducing a range of mysterious...mental states' (Heath 2008, p.66).

[11] It is worth noting that the scales of $u(o)$ and $u(a)$ range from 0 to 10, but this need not be so. The values only represent the relative value of outcomes (or actions). However, as Heath sees it, if the scale for $u(a)$ ranged from 0 to 100 and the scale for $u(o)$ ranged from 0 to 10, then it might suggest that the agent considers 'doing the morally right thing' to be ten times more important than 'obtaining the most satisfying outcome.'

[12] Even though Equation 2 captures his idea of the role of deontic norms, his Equations on pages 79 and 90 are more sophisticated developments, formulated to address particular problems.

[13] This is the agent's subjective conception of social harm, and is not derived from an objective social welfare function. Note also that omissions of activities, while seemingly equaling activities not done, may be recast to constitute an (other) activity done.

[14] Rabin assumes that if $y = q$, then the activity is done, as $g(q) = 0$. Also note that it is possible for utility to be less than zero if the estimated probability of social harm is high enough, but this is not problematic for the model.

[15] It is worth noting that even though both agents will decide $x = 1$ when $q < v = y$, there are different levels of utility between the agents: *R*-agents have more ($v - 0$) than *P*-agents ($v - q$, unless $q = 0$, in which case their utility is equal). Likewise, when both agents decide $x = 0$, *R*-agents will have less utility ($v - 1$) than *P*-agents ($v - q$, unless $q = 1$, in which case their utility is equal). In other words, the *R*-agent's utility is more sensitive to changes to their beliefs about q.

[16] Rabin (1995,p.9) states that the equality of y and v for *R*-agents can be understood as a moral-rule, 'don't do anything that is not defensible from the perspective of social welfare.'

[17] As they state, 'we will mostly refer to 'acts' and 'actors,' but these terms should be understood as referring to both legal rule-making and particular decisions. In fact, however, much of the analysis is relevant to moral (rather than legal) questions and to individual (rather than public) choices as well' (Zamir and Medina 2010, p.80).

[18] On several instances, Zamir and Medina state that in some cases the threshold function need not be a two-step process; rather, it is described as establishing a choice among alternative actions and a possible adjustment to the value of K. This is not explained very well until examples are given on pages 149-50 and 155-56, where the probability of outcomes from alternative actions can be compared.

[19] As with the other works reviewed here, the contribution of Rose to our purposes is not the objective of his work; while Equation 9 depicts the possible role of moral principles on individual behavior, this is not sufficient for his greater task. He does proceed to define another form of moral conduct, duty-based moral restraint.

However, he does not provide a formal representation of such a morality (Rose 2011, pp.114-34).

[20] However, this is a restriction that in some cases may limit our ability to represent non-consequentialist moral norms. See Suzumura and Xu (2001, 2003, 2009) for a more general approach to the choice variable in an attempt to delineate consequentialist and nonconsequentialist moralities.

[21] Though not mentioned in the coverage of Zamir and Medina (2010), they do consider the role of uncertainty and probabilities in moral decision making. Their incorporation into the model is straightforward, and most resembles that of Rose (2011).

[22] I will proceed to talk about the expected utility as if it is a Von Neumann-Morgenstern expected utility function, but this need not be the assumption for this formulation.

References

Anderson, Elizabeth (1987) *Value in Ethics and Economics*, Cambridge, MA, USA: Harvard University Press.

Ben-Ner, Avner and Louis Putterman (1998) 'Values and institutions in economic analysis', in Ben-Ner, Avner and Putterman, Louis (eds.), *Economics, Values, and Organization*, New York, NY, USA: Cambridge University Press, pp. 3-69.

Bicchieri, Cristina (2005), *The Grammar of Society*, Cambridge, MA, USA: Cambridge University Press.

Broome, John (1992), 'Deontology and economics', *Economics and Philosophy*, **8** (2), 269-88.

Bowles, Samuel (2003), *Microeconomics: Behavior, Institutions, and Evolution*, Princeton, NJ, USA: Princeton University Press.

Bowles, Samuel, and Sandra Polania-Reyes (2012), 'Economic incentives and social preferences: substitutes or complements?', *Journal of Economic Literature,* **50** (2), 368-425.

Dowell, Richard S., Robert S. Goldfarb, and William B. Griffith (1998) 'Economic man as a moral individual', *Economic Inquiry*, **36** (4), October 1998, 645-53.

Etzioni, Amitai (1986), 'The case for a multiple-utility conception', *Economics and Philosophy,* **2**, 159-83.

Etzioni, Amitai (1988), *The Moral Dimension*. New York, NY, USA: Free Press.

Falk, Armin, and Urs Fischbacher (2005), 'Modeling strong reciprocity' in Herbert Gintis, Samuel Bowles, Robert Boyd, and Ernst Fehr (eds.), *Moral Sentiments and Material Interests*, Cambridge, MA, USA: MIT Press, pp. 193-214.

Fehr, Ernst, and Klaus M. Schmidt (1999), 'A theory of fairness, competition, and cooperation', *Quarterly Journal of Economics,* **114**, 817-68.

Goldfarb, Robert S. and William B. Griffith (1991a), 'Amending the economist's 'rational egoist' model to include moral values and norms, Part 1: the problem' in Koford, Kenneth. J., and Miller, Jeffery. B. (eds.), *Social Norms and Economic Institutions*, Ann Arbor, MI, USA: University of Michigan Press, pp. 39-57.

Goldfarb, Robert S. and William B. Griffith (1991b), 'Amending the economist's 'rational egoist' model to include moral values and norms, Part 2: alternative aolutions' in Koford, Kenneth. J., and Miller, Jeffery. B. (eds.), *Social Norms and Economic Institutions*, Ann Arbor, MI, USA: University of Michigan Press, pp. 59-84.

Hammond, Peter (1986), 'Consequentialist social norms for public decisions' in Walter P. Heller, Ross M. Starr, and David A. Starett (eds.), *Social Choice and Public Decision Making: Essays in Honor of Kenneth Arrow*, vol. 1, Cambridge, MA, USA: Cambridge University Press, pp. 3-27.

Hammond, Peter (1988), 'Consequentialist foundations for modern utility', *Theory and Decision,* **25** (1), 25-78.

Heath, Joseph (2008), *Following the Rules,* New York, NY, USA: Oxford University Press.

Koford, Kenneth J., and Jeffery B. Miller (1991), 'Introduction' in Koford, Kenneth J., and Miller, Jeffery B. (eds.), *Social Norms and Economic Institutions*, Ann Arbor, MI, USA: University of Michigan Press, pp. 1-20.

Kuran, Timur (1998), 'Moral overload and its alleviation' in Ben-Ner, Avner and Putterman, Louis (eds.), *Economics, Values, and Organization*, New York, NY, USA: Cambridge University Press, pp. 231-66.

Rabin, Mathew (1995), 'Moral preferences, moral constraints, and self-serving Bias', Working Paper, Dept. of Economics, University of California, Berkeley, August 1995.

Rose, David C. (2011), *The Moral Foundation of Economic Behavior*, Oxford, UK: Oxford University Press.

Sen, Amartya (1973), 'Behavior and the concept of preference', *Economica*, **40** (159), 241-59.

Sen, Amartya (1974), 'Choice, Orderings, and Morality' in Stephan Korner (ed.), *Practical Reason*, New Haven, CT, USA: Yale University Press, pp. 54-66.

Sen, Amartya (1977), 'Rational fools: a critique of the behavioral foundations of economic theory', *Philosophy and Public Affairs*, **6** (4), 317-44.

Sen, Amartya (1982), 'Rights and agency', *Philosophy and Public Affairs*, **11** (1), 3-39.

Sen, Amartya (1987), *On Ethics and Economics*, Oxford, UK: Basil Blackwell.

Sen, Amartya (1997), 'Maximization and the act of choice', *Econometrica,* **65** (4), 745-79.

Sen, Amartya (1998) 'Foreword' in Ben-Ner, Avner and Putterman, Louis (eds.), *Economics, Values, and Organization*, New York, NY, USA: Cambridge University Press, pp. vii-xiii.

Suzumura, Kotaro, and Yongsheng Xu (2001), 'Characterizations of consequentialism and non-consequentialism', *Journal of Economic Theory*, **101** (2), 423-36.

Suzumura, Kotaro, & Yongsheng Xu (2003) 'Consequences, Opportunities, and Generalized Consequentialism and Non-Consequentialism', *Journal of Economic Theory*, **111** (2), 293-304.

Suzumura, Kotaro, & Yongsheng Xu (2009), 'Consequentialism and non-consequentialism: the axiomatic approach' in Paul Anand, Prasanta Pattanaik,

and Celmens Puppe (eds.), *The Handbook of Rational and Social Choice*, New York, NY, USA: Oxford University Press, pp. 346-73.

Vanberg, Victor (1988), *Morality and Economics: De Moribus Est Disputandum*. Bowling Green, Ohio, USA: Transaction Books.

White, Mark (2011), *Kantian Ethics and Economics*, Stanford, CA, USA: Stanford University Press.

Zamir, Eyal and Barak Medina (2010), *Law, Economics, and Morality*. Oxford, UK: Oxford University Press.

Ross A. Tippit is Assistant Professor in the Department of Social Sciences, Borough of Manhattan Community College-The City University of New York BMCC-CUNY, New York (USA) (rtippit@bmcc.cuny.edu).