Evaluating Allocations of Freedom

Itai Sher*
University of Minnesota

February 11, 2015

Abstract

This paper develops a formal approach to evaluating freedom in interactive settings based on the literatures on preference for flexibility and measurement of diversity. The approach posits that freedom has an instrumental component–grounded in preferences– and an intrinsic component. The philosophical justification and implications of the approach are considered. In particular, we discuss the nature of the required value judgments. Potential conflicts between freedom and efficiency are explored. On a technical level, the paper extends the notion of a *diversity measure* (Nehring and Puppe 2002) to menus of lotteries, which is what is needed to evaluate freedom when many agents seek flexibility simultaneously.

1 Introduction

While economists have had a deep and longstanding interest in freedom, with attempts to formalize this notion within social choice, it is fair to say that the evaluation of social institutions in terms of the freedom they provide has not yet been systematically integrated into mainstream economics and no consensus exists about how to think about the issue formally.[1] Greater consensus on this could potentially place freedom on a more equal footing with standard economic notions such as efficiency and welfare and so facilitate a discussion among economists about trade-offs between freedom and other social values. The aim of this paper is to take a step toward a workable formulation of a model of freedom that can be applied in the context of economic models.

A central problem when thinking about arrangements that promote freedom is that the exercise of freedoms on the part of some may restrict the freedoms of others. Bentham (1782) wrote, "You and your neighbor, suppose, are at variance: he has bound you hand

*I am very grateful to Martin van Hees for detailed comments on an earlier draft. I benefited from comments by Eric Maskin, Hendrik Rommeswinkel, Christopher Chambers, Peter Sher and participants at the Conference on Normative Ethics and Welfare Economics at the Becker Friedman Institute at the University of Chicago and seminar participants at UCSD.

[1]For a collection of philosophical views on freedom, see Carter, Kramer and Steiner (2007).

and foot, or has fastened you to a tree ... it is on account of what you have been made to suffer by the operation which deprives you of [your liberty] that the legislator steps in and takes an active part ... He must either command or prohibit ... he therefore cuts off on one side or the other a portion of the subject's liberty." For this reason, I construct an *interactive* model, in which many agents exercise freedoms simultaneously.

To evaluate freedom in an interactive setting, it is useful to first form a conception of freedom inherent in a decision facing a single individual. Barbera, Bossert and Pattanaik (2004) and Dowding and van Hees (2009) survey formal decision theoretic measures of freedom.[2] Another important approach to freedom in economics that deals with both formal and philosophical aspects of freedom is the capability approach (Sen 1979, Sen 1999, Nussbaum 2011)

1.1 The Instrumental Value of Freedom

One approach allied to economic modeling treats freedom instrumentally. In an individual decision problem, the value of a menu–a set of options–is conceptualized as the instrumental value of having access to the menu. To make this interesting and realistic, suppose that when the menu is evaluated, the agent does not yet have all of the information that will be relevant at the later time when she will actually choose from the menu. Clearly one important reason we would prefer to be free to choose rather than to be given the option that someone else predicts we will want is that freedom allows us to adjust our choices to the precise circumstances that obtain at the time of choice. The formal development of this idea in terms of a *preference for flexibility* draws on models that studied such preferences without explicitly invoking the concept of political freedom (Koopmans 1964, Goldman 1974, Kreps 1979, Nehring 1999, Dekel, Lipman and Rustichini 2001). Arrow (1995) proposed that such a model could be used as a model of freedom.

1.2 From Choices to Interactions

It is relatively straightforward, then, to define the value of a menu as its instrumental value, but how does one define the value of freedom in a social interaction? A natural extension would be to define the value of freedom to be the instrumental value of participating in the interaction. This proposed definition must confront a circularity: Freedom and choice mutually determine one another: My freedom is (partially) determined by your choices; my freedom also influences my choices, which in turn determines your choices. This is the common sort of circularity that is found in game theory. It is also the specific circularity discussed by Bentham in the quotation above.

[2]This literature includes Jones and Sugden (1982), Pattanaik and Xu (1990), Klemisch-Ahlert (1993), and Foster (2010), among many others.

One way of cutting through the circularity is in the spirit of game theory. To simplify matters, a social interaction consists in the rules that govern the interactions–laws, sanctions, physical and technological constraints, etc.–the internal states of the agents, including preferences and beliefs, and the resulting behavior. Behavior (and beliefs) are determined in part by another circularity: the agents' predictions about others' behavior. Now given the rules and internal states, consider the behavior that will actually emerge. Consider some specific agent, say Ann. Call Ann's **equilibrium menu** the set of outcomes that Ann can achieve holding *fixed* the behavior of all agents other than Ann. (One can also call this Ann's *effective* menu if one does not want to take on all the connotations of equilibrium.) We can now apply the above understanding of the freedom inherent in menus to the equilibrium menu to derive Ann's freedom in the social interaction. One insight that immediately emerges from this approach is that it is not just the rules governing a given arrangement that determine an agents' freedom; that freedom is also determined by whatever behavior ends up prevailing.[3]

To illustrate the ideas with an example, suppose I want to compare my freedom in a situation where I own a bicycle to my freedom in a situation where the bicycle is communally owned. The bicycle, were I to own it, grants me freedom to go to various locations. I may think of the bicycle as the menu of locations to which it gives me access. But suppose the bicycle is communally owned. Then when I go to retrieve it, it may not be present. Suppose that the probability that the bicycle is in use is 1/3. Then under the communal arrangement, we may formally associate the bicycle with the menu that contains for each location L a lottery that allows me to go to L with probability 2/3 and keeps me at home with probability 1/3. The menu also contains the alternative corresponding to the choice of staying home with certainty.

It may be that the probability that someone will actually want to use the bicycle at

[3]This point is somewhat related to a distinction emphasized by Sen (2009): "One example of some interest and relevance is an important distinction between two different concepts of justice in early Indian jurisprudence - between *niti* and *nyaya*. The former idea, that of *niti*, relates to organizational propriety as well as behavioural correctness, whereas the latter, *nyaya*, is concerned with what emerges and how, and in particular the lives that people are actually able to lead." While *niti* and *nyaya* relate directly to justice rather than freedom, one can make a somewhat analogous distinction between the freedom inherent in the formal rules governing social interaction (incorporating also such hard factors as physical constraints) and the actual freedom that emerges when behavior is factored in. Van Hees (2000) calls freedom that depends on the law *legal freedom*. The conception of freedom studied here depends essentially on not only the rules, but the resulting behavior. Other formal perspectives make the measure of freedom depend only on the rules. This is true if one takes the construct of an *effectivity function* as a *measure* of freedom, or power (Moulin and Peleg 1982, Peleg 1984, Abdou and Keiding 1991, Peleg and Peters 2010). That is not to say that this literature is not concerned with behavior; a great deal of analysis studies the relation between effectivity functions and strategic behavior. For a related game-theoretic approach using rights structures to study the compatibility of freedom and efficiency, see Van Hees (1999). An approach to measuring freedom in games based on mutual information, which does depend on strategic choices and not just rules, has recently been developed by Rommeswinkel (2014). Using a game-theoretic analysis, Dowding and van Hees (2007) show how changes in the other agents' preferences, and the resulting changes in their behavior, can impact an agent's freedom. Sections 3.8-3.9 of Dasgupta (1995) include a general framework that is similar to the the one described in this subsection.

the same time that I want to use it is 1/6 rather than 1/3. The reason that the bicycle is missing more often is that when someone thinks that he may want to use the bicycle in the near future, he stores it in his room to ensure it is available when he wants it.[4] This illustrates that in order to calculate my effective freedom, we must know how the rules of the game will cause others to behave, which depends in part on their view of how I will behave (How likely am I to take the bicycle at the time they want to use it?), which in turn depends on my forecast of their behavior, and so on. Thinking about a broader array of communally shared items, we see that the amount of freedom that a community menu can support depends on the preferences and behavior of individuals. If everyone always wants to use the same items at once, the community menu can support very little freedom, but if it is unlikely that two people want to use the same item at once, then it may be almost as if each agent had a duplicate of the communal menu as their own. So effective freedom does not just depend on the social arrangement–common ownership–but on preferences and prevailing behavior.

1.3 The Intrinsic Value of Freedom

1.3.1 Motivation

The above construction is based on a view of the value of freedom as instrumental. Philosophers have long debated whether freedom has only instrumental or whether freedom also has intrinsic value. Sen (1999), for example, writes, "In general, the case for seeing intrinsic importance in autonomy and liberty is not easy to escape, and this can easily conflict with no-nonsense maximization of the utility consequences."[5] The following example illustrates the sort of considerations that might drive some to view freedom as having intrinsic value. Consider the freedom to criticize the government. Suppose it is well known that Ann would never criticize the government because she feels that she should be loyal to the government no matter what. There is then no instrumental value to Ann in being able to criticize the government, as this option will never be exercised. Would it then be acceptable for the government to eliminate this option for Ann? The feeling that the answer is no might drive one to view the value of this freedom as being at least in part intrinsic, and not purely instrumental. For further arguments and counter-arguments on this issue, see Section 3.4.

[4] For expositional simplicity, I have not formally represented this option in the description of the menu above.

[5] In mentioning autonomy, Sen is alluding to a positive conception of liberty that I do not explicitly pursue in this paper. For the distinction between positive and negative freedom, see Berlin (1959).

1.3.2 Modeling Intrinsic Value

To construct a formalism that can accommodate *both* the instrumental and the intrinsic values of freedom, I build on the theory of *diversity measures* of Nehring and Puppe (2002).[6] Diversity measures have been used to study the diversity inherent in a set of objects (or species), and also the freedom inherent in a choice set (Nehring and Puppe 2008) as how much freedom a choice set grants to an individual is related to how diverse the alternatives are.

To be more precise, consider a menu from which an agent may choose. We would like to evaluate the freedom inherent in such a choice set. Suppose that there is some set of attributes that are relevant to the decision. For example if a menu consists of transportation options, the relevant attributes might be that the mode of transportation is fast, that it is comfortable, that it is inexpensive, and so on; one may also include conjunctive attributes such as that it is both fast and inexpensive. Each attribute is assigned a value. Then the value of the menu is the sum of the values of the attributes that are instantiated by some alternative in the menu.[7] The value of a menu is then understood as the value of being able to bring about various attributes that the menu allows. It is possible to encode many specific cases which differ qualitatively within this general framework.

Nehring (1999) and Nehring and Puppe (2002) have shown that ranking choice sets in terms of diversity is formally equivalent to ranking choice sets in terms of preference for flexibility. However, while formally identical, the diversity and flexibility interpretations differ substantively. In particular, the information that one needs if one takes the diversity measure interpretation seriously differs from the information one needs if one takes the preference for flexibility interpretation seriously. What one needs to construct the diversity measure is a value judgment about the importance to the agent's freedom of being able to bring it about that the outcome has various attributes. What one needs to construct a ranking based on preference for flexibility is knowledge about the preferences of an individual or a population of individuals.

I extend the notion of a diversity measure to sets of lotteries because that is what is needed in an interactive context to deal with outcomes generated when agents use mixed strategies. (Nehring (1999) and Nehring and Puppe (2002) deal with the case of lotteries over menus, but what is needed here is an analysis of menus of lotteries.) I show that under the extension to sets of lotteries the equivalence of the diversity measure and preference for flexibility approaches continues to hold (Proposition 2).

[6] This theory generalizes that of Weitzman (1992) and Weitzman (1998).

[7] For the interpretation of the value of conjunctive attributes, see the discussion of the attribute *oppose-not-punished* in Section 2.1.

1.4 Integrating Intrinsic and Instrumental Value

I use the mathematically equivalent notions of (i) preference for flexibility and (ii) diversity measures to construct a hybrid approach to evaluating freedoms both on the basis of their intrinsic and instrumental values. Suppose that we have two value functions, ν^* and μ°, measuring, respectively the *instrumental value* $\nu^*(M)$ and the *intrinsic value* $\nu^\circ(M)$ of a menu M. $\nu^*(M)$ gives the *option value* of menu M for an agent who does not yet know exactly what she will prefer to choose as described above in Section 1.1. $\nu^\circ(M)$ gives the *access value* of menu M, that is the value of menu M in virtue of providing access to options with certain intrinsically valuable attributes. The value of access to these attributes is not derived from the utility realized in circumstances that the agent will realize them; it is the access itself that is valuable.[8,9] I prove that if there is an overall value ν function that has the same mathematical form as both ν^* and ν° (specifically it satisfies (5) below), and ν responds positively to increases in both intrinsic and and instrumental value (and nothing else), then ν is a weighted average of ν^* and ν°; that is, there exists a weight α, between 0 and 1, such that for all menus M, $\nu(M) = \alpha\nu^*(M) + (1-\alpha)\nu^\circ(M)$ (see Propositon 3). I call ν the **hybrid** measure. So freedom has an instrumental component ν^* and an intrinsic component ν°. That alternatives contribute to access value over and above their option value helps to capture the idea that the value of the freedom to take an action is distinct from the value of that action (see, e.g., Kramer (2003)).[10] [11]

Mathematically, the hybrid measure ν can be generated by a diversity measure – distinct from the diversity measure that generates ν° – that treats instrumentally valuable attributes as if they were intrinsically valuable. Symmetrically, by the equivalence of the diversity and flexibility approaches, the resulting value function ν is that of some hypothetical agent with

[8] This does leave open the possibility that the access value depends in some way on the characteristics of the individual. For example the value of being able to make a certain utterance might depend on how the agent understands the utterance.

[9] The distinction between the option value and access value of freedom is reminiscent of the distinction between freedom as an *exercise* concept and as an *opportunity* concept (Taylor 1979). The terms option value and access value are better adapted to formal machinery, and opportunity and exercise value have additional connotations that are not implied. For instance, Taylor associates exercise freedom with authenticity and self-realization.

[10] On the instrumental flexibility account, the value of an alternative depends on the menu to which it is added: For example, the instrumental value of adding the option to see a particular movie m to a menu containing many movies is likely to be less than the value of adding that movie to a menu containing no other movies. So even on the instrumental account, the contribution of an alternative to the value of an agent's freedom differs from that alternative's stand-alone value.

[11] Carter (1999) uses similar considerations to distinguish the *specific* and *non-specific* values of freedom. The distinction between specific and non-specific value of freedom is distinct from the distinction between intrinsic and instrumental value (see Chapter 2 of Carter (1999)). On the basis of considerations that can be articulated in part using formal tools similar to the ones employed here, I am somewhat skeptical of the specific/non-specific dichotomy, but expanding such an argument, as well as the complementary task of presenting counterargument to Carter's negative conclusions against value-based measures of freedom, would require too much space to be made here. There is also the distinction between the *quantity* of freedom and its *value*. The utility of this distinction for evaluating freedom must also be discussed elsewhere.

preference for flexibility; the implicit distribution of preferences of the hypothetical agent do not coincide with the actual distribution of instrumental payoffs – as represented by ν^* – but is altered so as to give value to hypothetical choices that are important from the standpoint of freedom but may not be exercised.

The proof of the proposition establishing the hybrid measure is a variant of the proof of Harsanyi's aggregation theorem (Harsanyi 1955). The contrast in interpretation is illuminating: Harsanyi's theorem established that an expected utility maximizer whose utility responds positively to that of expected utility maximizers in the population should be a utilitarian. In contrast, in light of the flexibility representation of the hybrid measure, Proposition 3 establishes conditions under which a benevolent social judge who takes intrinsic freedom seriously should choose among institutions by treating intrinsically valuable options *as if* they entered into agents' utility functions.

1.4.1 Extending The Hybrid Measure to Interactions

The hybrid approach can be applied in an interactive setting in a manner completely parallel to the instrumental approach. Again look at matters from the point of view of a single agent, say Ann. The rules of the game and the behavior of all agents other than Ann determine an equilibrium menu for Ann in the same manner as above. The difference is that we *evaluate* the equilibrium menu via the hybrid rather than the instrumental measure.

Even in the limiting case where freedom is *only* intrinsically valuable (all weight is put on the intrinsic value function), the consequences of instituting a social arrangement are important to its evaluation. This is because the consequences determine the menu from which an agent actually chooses. Even if this menu is evaluated intrinsically, what menu the agent faces depends on the consequences. For illustration, return to the Bentham quotation of the opening paragraph. Suppose that the legislator does nothing, and consequentially, my neighbor ties me to a tree. Even if the loss of my freedom is only to be judged as intrinsically bad, the fact that I lost my freedom depends on what my neighbor does in virtue of not being bound by the legislator. This feature of the theory–the *indirect* role of consequences even for intrinsic evaluation–is not a necessary feature of any theory. One could imagine instead a *purely* procedural theory that evaluated institutions without any reference to their consequences. The indirect role of behavioral consequences,while initially non-obvious–especially when valuing freedom intrinsically–is, upon reflection, highly plausible.

Proposition 4 extends Proposition 3 from single-agent decisions to social interactions. The social objective can be interpreted as weighing instrumental benefits against costs associated with restrictions on intrinsically valuable freedoms. Thus consequentialist benefits are traded off against violations of deontic constraints.[12]

[12]For a related approach, see Zamir and Medina (2008).

1.4.2 The Required Value Judgments

What is the sum total of value judgments required by the above scheme? Value judgments are required to spell out both (i) the nature and value of the attributes that are intrinsically valuable with respect to freedom, and (ii) the trade-off between instrumental and intrinsic values. It is important not to forget that the aggregation involved even in making instrumental judgments of value depend on value judgements concerning how the interests or internal states of different agents should be aggregated,[13] and perhaps even about the relative importance of different internal states of a single individual.[14] One goal of the paper is to clarify the structure of the mix of value judgments and factual knowledge required for making judgments about freedom.

1.5 Consequences

Important normative consequences of the approach include:

Conflict with Pareto. In that the hybrid social objective takes into account liberties that are not reducible to individual preferences, it can conflict with the Pareto criterion. This issue, which relates to Sen's (1970) liberal paradox,[15] is illustrated in the context of a detailed example in Section 6 and discussed in Section 7.4.

Evaluation of Procedural Aspects of Mechanisms. A key aspect of the hybrid approach is that it evaluates not just consequences – as in standard economic models – but also the *procedural* aspects of mechanisms. In particular, the approach evaluates differently a situation in which an agent chooses an alternative and a situation in which that alternative is imposed (see Sections 4 and 6 for illustrations).

Cost-Benefit Analysis. Incorporating an evaluation of freedom could have important effects on cost-benefit analysis (see Section 4 for an application to the value of life). The current paper provides a framework for thinking rigorously about incorporating considerations of liberty in such analyses. (For related frameworks incorporating other moral values,

[13] Even the utilitarian who has some way of measuring happiness, where the measurement has cardinal significance that is interpersonally comparable must make the value judgment that this measurable quantity is what matters normatively.

[14] The need for value judgments if one wants to assess the the interests of a single individual is most obvious when the individual is "irrational" in some way; for example, consider a smoker who is dynamically inconsistent, and would like to bind his hands to prevent himself from smoking. If we side with the person's desire to quit smoking, we are making a value judgment even if this is supported by our understanding of the nature of addiction. (For a recent literature that attempts to extend traditional notions of welfare economics to such irrational or inconsistent agents, see Green and Hojman (2007), Bernheim and Rangel (2007), Bernheim and Rangel (2009), and Rubinstein and Salant (2012); see also Thaler and Sunstein (2003) and Thaler and Sunstein (2008).) However, even if an agent is rational in a traditional economic sense, a judgment is required–however obvious it we may think it is–that the individual's judgement about what is in his interest is the judgment that we should adopt. The thesis that an individual is sovereign with respect to his own good is a substantive ethical thesis, even if it is correct. Indeed, some philosophers have denied it.

[15] In the social choice framework, Gibbard (1974) formalizes the intuitive proposition that rights can conflict all on their own.

see Zamir and Medina (2008) and Lowry and Peterson (2011).)

1.6 Outline

The outline of the paper is as follows. Section 2 develops the theory of diversity measures, and specifically the theory of *stochastic diversity measures* as applied to menus of lotteries. Section 3 establishes the formal equivalence of the diversity and flexibility approach and introduces and justifies the hybrid evaluation of freedom in terms of both instrumental and intrinsic values. Section 4 illustrates the hybrid approach with an example on the value of life. Section 5 extends the framework to an interactive setting. Section 6 presents a detailed application of the interactive framework to an example concerning control of the personal sphere. Section 7 discusses various philosophical issues raised by the framework. An appendix presents proofs omitted from the main text.

This paper contains a mix a technical and philosophical arguments. I have attempted to include informal explanations in the technical sections to make them accessible to a broader audience. Much of the discussion, such as the argument for intrinsic values of Section 3.4 and the philosophical discussion of Section 7 should be accessible to a nontechnical audience.

2 Freedom Measures

To evaluate freedom *intrinsically*, I employ a notion originally applied to measure the diversity of choice sets (Nehring and Puppe 2002).[16] The intuitive idea that this model attempts to capture is that there is an intrinsic value to having access to alternatives with certain freedom-relevant attributes.

Because the purpose of this section is to develop the formal theory and illustrate it, I will not provide arguments that the specific examples used are of intrinsic rather than instrumental value. The question of which attributes really have intrinsic value is not part of the formal calculus for reasoning about freedom, but is rather a substantive question preliminary to the application of the formal calculus. It is also important to bear in mind that the same attribute may have both instrumental an intrinsic value (see Section 3.5).

For further a discussion of diversity measures beyond their applications to freedom as well as of their broad expressive power to model many different kinds of diversity, the reader is referred to Nehring and Puppe (2002).

[16] Freedom and diversity are intuitively related insofar as more diverse choice sets offer more freedom (Nehring and Puppe 2008). For this equation of diversity and freedom to hold, it must be that the alternatives are diverse with respect to attributes that are significant to choice.

2.1 Deterministic Freedom Measures

Let Z be a set of outcomes. A subset of Z may be viewed in two ways: It may be viewed as a **menu** from which a choice is made, or it may be viewed as an **attribute** shared by elements of that set. The idea behind a diversity measure, which we refer to as a **freedom measure** to emphasize the application at hand, is to assign a value to each attribute and then to value each menu as the sum of values of its attributes.

Specifically, if $\Lambda(A)$ is the value of attribute A, then the value of menu M, which we denote by $\nu(M)$ is given by:[17]

$$\nu(M) = \sum_{A \subseteq Z : M \cap A \neq \emptyset} \Lambda(A) \tag{1}$$

The interpretation is that for each attribute, there is a value assigned to the freedom associated with the ability to bring about that attribute. The value of a menu is then the sum of values associated with abilities allowed by the menu to bring about attributes.

To illustrate, suppose the freedom-relevant attributes are *supporting the government* and *opposing the government*, abbreviated *support* and *oppose*, respectively. (For the moment, assume all other attributes A are given no value: $\Lambda(A) = 0$; I add other valuable attributes below.) Many outcomes z may share an attribute: Outcome z_1, in which *the agent attends a pro-government rally*, and z_2, in which *the agent writes a newspaper editorial praising the government*, are instances of the attribute *support*. Outcome z_3 in which *the agent attends an anti-government rally and is not punished for doing so* is an instance of *oppose*. Outcome z_4 in which *the agent goes for a walk and is not punished* is an instance of neither attribute. Menu $M_1 = \{z_1, z_3\}$ instantiates (i.e, *intersects*) both the attribute *support* and the attribute *oppose*, and so has value $\Lambda(\textit{support}) + \Lambda(\textit{oppose})$. $M_2 = \{z_1, z_2, z_4\}$ and $M_3 = \{z_1, z_4\}$ both instantiate only *support* and so both attain value $\Lambda(\textit{support})$, allowing less valuable freedom than M_1. To differentiate M_2 and M_3 in terms of freedom-value, we would have to introduce a freedom relevant attribute – say, *writing an editorial about the government* – which is instantiated by z_2, but not by z_1 and z_4. Then z_2 would instantiate two freedom-relevant attributes, *support* and *editorial*, and contribute freedom value $\Lambda(\textit{support}) + \Lambda(\textit{editorial})$ to the menus to which it belongs. Let z_5 be the outcome in which *the agent protests the government and is punished for doing so.* Let *not-punished* and *oppose-not-punished* be the attributes of *not being punished* and of *opposing the government while not being punished.* Then the value of $\{z_4, z_5\}$ is $\Lambda(\textit{oppose}) + \Lambda(\textit{not-punished})$, which is less than the value of $\{z_2\}$, which is $\Lambda(\textit{oppose}) + \Lambda(\textit{not-punished}) + \Lambda(\textit{oppose-not-punished})$, capturing the intuition that it is more valuable to be able to simultaneously oppose and not be punished than it is to achieve these outcomes separately. Here $\Lambda(\textit{oppose-not-punished})$ should be interpreted as the *incremental* value of being able to realize the attributes *op-*

[17] Assume that $\Lambda(A) \geq 0, \forall A \subseteq Z$.

pose and *not-punished* together as opposed to separately. As all the examples illustrate, this model interprets the value of freedom as the value of *having access* to alternatives instantiating certain freedom-relevant attributes.

In the interactive setting we study, we will often have to evaluate menus that consist not merely of alternatives in Z, but rather of lotteries over Z. For this reason, we must extend the notion of a diversity or freedom measure to cover menus of lotteries,[18] which is what we now proceed to do. For this reason we distinguish a **deterministic freedom measure**, as defined above, from a **stochastic freedom measure**, as we define below.

2.2 Stochastic Freedom Measures

Let Z be a finite set of outcomes. $\Delta(Z)$ is the set of lotteries on Z. Each lottery in $\Delta(Z)$ cab be represented as a vector $\beta = (\beta_z : z \in Z)$ where β_z is the probability of outcome z according to β.

2.2.1 Attributes

One could define a *stochastic attribute* – an attribute of lotteries – as an arbitrary set of lotteries. This would impose essentially no structure on stochastic attributes. Instead, I impose structure that allows an intuitive grasp of attributes and the relations between them.

Imagine that outcomes specify many dimensions of life: income, geographical location, health, etc. Let $A(care, \geq q)$ be the attribute: *receiving medical care with probability at least q*. In contrast to the attribute $A(care, = q)$ of *receiving medical care with probability exactly q*, the typical lottery β satisfying $A(care, \geq q)$ still satisfies $A(care, \geq q)$ if the probabilities in β are perturbed slightly. So attributes of the from $A(care, \geq q)$ are less fragile than those of the from $A(care, = q)$. This justifies a focus on the former more robust type of attribute. The stochastic attribute $A(care, \geq q)$ can be decomposed into two components: the attribute "type", which is a deterministic attribute – in this case, *receiving medical care* – and a "degree", which is q. Thus $A(care, \geq .5)$ and $A(care, \geq .75)$ are two attributes of the same type but of different degrees. In contrast $A(children, \geq .5)$, which is the attribute of *having children with at least probability .5*, is of a different type than $A(care, \geq .5)$.

A natural generalization of attributes like $A(care, \geq q)$ is: *the weighted average of the probabilities that several kinds of medical care are available is at least q*, where particular weights are assigned to particular kinds of medical care. Other similar examples might involve taking weighted averages of probabilities of access to different leisure activities or ways of protesting the government. In such cases, the "type" of the attribute is not a single deterministic attribute, but a particular "weighted average" of deterministic attributes.

[18] The analysis of Nehring (1999) and Nehring and Puppe (2002) applies to lotteries over menus, but what we need is to be able to analyze menus of lotteries.

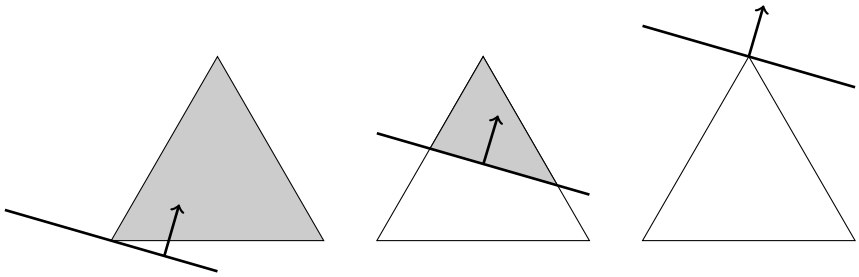

Figure 1: Stochastic Attributes

The above discussion suggests that we construct stochastic attributes by assigning weights to deterministic attributes. As explained below, assigning weights to *outcomes* is equally general (within my model). Let $v = (v_z : z \in Z)$ be a weight vector assigning a weight v_z to each outcome z. I impose two normalizations on v: (i) $\sum_{z \in Z} v_z = 0$, and (ii) $\|v\| = 1$.[19] V is the set of weight vectors satisfying (i) and (ii). Weight vectors in V always assign negative weight to some states, but this is immaterial: I could have imposed different normalizations so that all weights would be nonnegative without altering the set of stochastic attributes to be defined below.

I consider stochastic attributes of the form

$$\{\beta \in \Delta(Z) : v \cdot \beta \geq h\} \tag{2}$$

where $v \cdot \beta = \sum_z v_z \beta_z$ and h is a real number. If $h \leq \min_{z \in Z} v_z$, then (2) is the set of all lotteries; if $h > \max_{z \in Z} v_z$, then (2) is the empty set. For $q \in [0, 1]$, let $h(v, q)$ be the number solving $q = \frac{h(v,q) - \min_{z \in Z} v_z}{\max_{z \in Z} v_z - \min_{z \in Z} v_z}$. Thus, $h(v, q)$ is the value of h in which $(q \times 100)\%$ of the distance from $\min_{z \in Z} v_z$ to $\max_{z \in Z} v_z$ has been traversed. Define the stochastic attribute $A(v, q)$ by $A(v, q) = \{\beta \in \Delta(Z) : v \cdot \beta \geq h(v, q)\}$.

Figure 1 illustrates $A(v, q)$ for a fixed v and three values of q. The vertices of a triangle correspond to three possible outcomes, and each point in the triangle represents a lottery, where the closer the point is to a vertex, the more weight the lottery puts on the corresponding outcome. The shaded region corresponds to the set of lotteries satisfying the stochastic attribute $A(v, q)$. v corresponds to the direction of the arrow – perpendicular to the line setting the boundary of $A(v, q)$ – pointing into the shaded region. q measures how demanding the attribute is: Increasing q corresponds to moving the line up and so reducing the portion of the triangle that is shaded. When $q = 0$, the attribute is not demanding at all, and the entire triangle is shaded (the left panel), when $q = 1$, then the attribute is maximally demanding so that (typically) only a single degenerate lottery putting probability

[19] (ii) $\|v\| = \sqrt{\sum_z v_z^2} = 1$ normalizes the length of v to be 1, so that v may be interpreted as a direction, and (i) then means that v can be interpreted as a direction within the simplex $\Delta(Z)$.

1 on a single outcome instantiates the attribute (the right panel).[20] When q is between 0 and 1, then $A(v,q)$ resembles the shaded region in the middle panel. Given this geometric interpretation, we refer to v as the **direction** – rather than the type – of the attribute, and q as the **degree** of the attribute.

We return now to the prototypical examples motivating consideration of this particular class of stochastic attributes. For any deterministic attribute A and lottery β, let $\beta(A)$ be the probability that the outcome instantiates attribute A. Then for any nonempty subset B of $\Delta(Z)$, there is a collection of deterministic attributes $A_1,\ldots,A_k$ and a collection of nonnegative weights $w_1,\ldots,w_k$ with $\sum_i w_i = 1$ and a number $p \in [0,1]$ such that $B = \{\beta \in \Delta(Z) : \sum_i w_i\beta(A_i) \geq p\}$ if and only if there exists $v \in V$ and $q \in [0,1]$ such that $B = A(v,q)$.[21] This shows that our model of attributes exactly captures the natural sort of attributes that motivated the analysis. Let n be the total number of possible outcomes and m be the number of outcomes in deterministic attribute A. Then in the special case where the stochastic attribute is *the probability of A is at least q* (i.e., the special case where $A_1,\ldots,A_k$ contains only one element), the stochastic attribute can be represented as $A(v,q)$ where $v_z = \sqrt{\frac{n-m}{nm}}$ if z is in A, and $v_z = -\sqrt{\frac{m}{n(n-m)}}$ if z is not in A.

Define the set of stochastic attributes to be $\mathcal{A} = \{A(v,q) : v \in V, q \in [0,1]\}$. Observe that for all v, $A(v,0) = \Delta(Z)$. When $0 < q < 1$, each stochastic attribute has a unique representation in terms of a pair (v,q).[22] The fact that the representation is not unique when $q = 0$ or $q = 1$ will not affect the way we evaluate attributes below. As in the case of deterministic attributes, some stochastic attributes are subsets of others. However, unlike deterministic attributes, the set of stochastic attributes is not closed under intersection. This will be discussed further in Section 2.3.

2.2.2 Measures

This section develops the notion of a stochastic freedom measure analogous to that of a deterministic freedom measure presented in Section 2.1. $\tilde{\lambda}(v,q)$ will represent the value of having access to a lottery with stochastic attribute $A(v,q)$ in a way analogous to that in which $\Lambda(A)$ represented the value of having access to an outcome with deterministic attribute A.

For any $v \in V$, attribute $A\left(v,\frac{2}{3}\right)$ is more difficult to satisfy than $A\left(v,\frac{1}{2}\right)$: a lottery β satisfying $A\left(v,\frac{2}{3}\right)$ also satisfies $A\left(v,\frac{1}{2}\right)$, but there will be lotteries β' satisfying $A\left(v,\frac{1}{2}\right)$ but not $A\left(v,\frac{2}{3}\right)$. The "sum" of attribute values of attributes satisfied by β will include both $\tilde{\lambda}\left(v,\frac{1}{2}\right)$ and $\tilde{\lambda}\left(v,\frac{2}{3}\right)$, whereas the corresponding sum for β' will include $\tilde{\lambda}\left(v,\frac{1}{2}\right)$ but not

[20] If v is perpendicular to one of the sides of the triangle, then $A(v,1)$ may be that side (if v points out of the triangle from that side).

[21] There is generally not a *unique* way of expressing an attribute $A(v,q)$ in the form $\{\beta \in \Delta(Z) : \sum_i w_i\beta(A_i) \geq p\}$.

[22] This follows from the normalizations defining V.

$\tilde{\lambda}\left(v, \frac{2}{3}\right)$. So $\tilde{\lambda}(v,q)$ should be interpreted as the marginal value associated with increasing the level q at which $A(v,q)$ is instantiated. This is analogous to the fact that in the theory of deterministic attributes, if A, B and $A \cap B$ are the only relevant attributes, the value $\Lambda(A \cap B)$ of the conjunctive attribute $A \cap B$ should really be interpreted as the incremental value accruing to a menu that instantiates the attributes A and B together within a single item, as opposed to a menu that only satisfies them separately in two different items.

I assume that $\tilde{\lambda}(v,q)$ depends only on the direction v of the attribute $A(v,q)$ and not on the degree q. So I write $\tilde{\lambda}(v,q) = \lambda(v)$. When $A(v,q)$ represents that *deterministic attribute A is realized with probability at least q*, this means that increasing the probability with which A is satisfied from q to $q+\epsilon$ achieves the same increase in value of intrinsic freedom as increasing this probability from q' to $q'+\epsilon$ for any other q'. When $A(v,q)$ is of the more general from, *the weighted average of probabilities of $A_1, \ldots, A_k$ is at least q*, the assumption has a similar interpretation. The *total* – as opposed to marginal – value of instantiating $A(v,q)$ – which includes the values of instantiating $A(v,q')$ for all $q' \leq q$ – is $\lambda(v)q$. The value of a menu M of lotteries is attained by integrating the marginal values of all stochastic attributes instantiated in M (see (3)), where this integral is analogous to the sum (1) in the case of deterministic freedom measures.

I now develop these ideas rigorously. A **stochastic freedom measure** is represented by a tuple (λ, μ) where μ is a measure on V and $\lambda : V \to \mathbb{R}_+$ is an integrable function with respect to μ. We interpret μ as a reference measure, which is intended to be chosen as a uniform measure on V or on some subset of V when possible, although this uniformity condition is not formally required. So the substance of the freedom measure is intended to be encoded in λ.[23] Define $\tilde{\lambda} : V \times [0,1] \to \mathbb{R}$ by $\tilde{\lambda}(v,q) := \lambda(v)$. Define the product measure $\tilde{\mu} := \mu \times \ell$ where ℓ is Lebesgue measure on $[0,1]$. $\mathcal{M}$ is the set of possible menus of lotteries. Formally, $\mathcal{M}$ is the set of non-empty closed subsets of $\Delta(Z)$.[24] For any menu M of lotteries, define the value of menu M, $\nu(M)$ by:[25]

$$\nu(M) = \int_{\mathcal{A}(M)} \tilde{\lambda} d\tilde{\mu}, \quad \text{where} \quad \mathcal{A}(M) = \{(v,q) : A(v,q) \cap M \neq \emptyset\}. \tag{3}$$

(3) is the stochastic analog of the deterministic (1). I write $\nu = \nu_{\lambda,\mu}$ when I want to express the dependence of ν on (λ, μ).

[23] Instead of (λ, μ), I could have specified a single measure τ, which relates to (λ, μ) via $\tau(E) = \int_E \lambda d\mu$. Two pairs (λ, μ) and (λ', μ') that lead to the same τ represent the same diversity measure. I choose to represent diversity measures via a pair (λ, μ) because this leads to a more intuitive presentation, which compensates for the lack of uniqueness.

[24] $\Delta(Z)$ is represented as $\left\{(\beta_z)_{z \in Z} \in \mathbb{R}^Z : \beta_z \geq 0, \forall z \in Z; \sum_{z \in Z} \beta_z = 1\right\}$, which is a subset of $\mathbb{R}^Z$.

[25] The set $\mathcal{A}(M)$ is measurable for all $M \in \mathcal{M}$.

2.3 Recovering a Deterministic Freedom Measure

From any stochastic freedom measure, one can recover a deterministic measure applying to deterministic menus. In particular, for outcome z, let δ_z be the degenerate lottery that selects outcome z with probability 1. Let $\mathcal{M}_d$ be the set of menus that contain only such degenerate lotteries. A menu $M \in \mathcal{M}_d$ is a **deterministic menu**. Similarly, let $\mathcal{A}_d$ be the set of nonempty subsets of $\bar{Z} := \{\delta_z : z \in Z\}$. So $\mathcal{A}_d$ is the set of **deterministic attributes**. Clearly a deterministic attribute is *not* a stochastic attribute, even in a degenerate sense. For any deterministic attribute $A \in \mathcal{A}_d$, define:

$$\Lambda_d(A) := \int_{\mathcal{L}(A)} \tilde{\lambda} d\tilde{\mu} \quad \text{where} \quad \mathcal{L}(A) = \left\{(v,q) : A(v,q) \cap \bar{Z} = A\right\}. \tag{4}$$

Below, I write $\Lambda_d^{(\lambda,\mu)}$ to express the dependence of Λ_d on (λ,μ). If we define the **deterministic projection** of a stochastic attribute $A(v,q)$ to be the set of "deterministic" lotteries $\delta_z \in \bar{Z}$ that satisfy the $A(v,q)$, then (4) defines the value of deterministic attribute A to be the "sum" of values of stochastic attributes whose deterministic projection is A. Thus, Λ_d is derived from (λ,μ) through a kind of marginalization.

Proposition 1 *For all deterministic menus M,*

$$\nu(M) = \sum_{A \subseteq \bar{Z} : M \cap A \neq \emptyset} \Lambda_d(A).$$

Moreover, for any deterministic freedom measure $\Lambda : 2^{\bar{Z}} \to \mathbb{R}_+$ with $\Lambda\left(\bar{Z}\right) = 0$,[26] *there exists a stochastic freedom measure (λ,μ) such that for all $A \subseteq Z$, $\Lambda_d^{(\lambda,\mu)}(A) = \Lambda(A)$.*

This shows that we can recover a deterministic freedom measure Λ_d from any stochastic freedom measure (λ,μ), and moreover the model of stochastic freedom measures is completely general in the sense that it puts no (substantive) constraint on the induced deterministic freedom measure. (As $\bar{Z}$ is an attribute common to all deterministic menus, setting $\Lambda(\bar{Z}) = 0$ does not alter the ranking of menus, and hence may be viewed as a normalization.) Note finally that despite the fact that set of *stochastic* attributes is not closed under intersection, it is possible that for deterministic attributes, A and B, $\Lambda_d(A) > 0, \Lambda_d(B) > 0$ and $\Lambda_d(A \cap B) > 0$.

[26] $2^{\bar{Z}}$ is the set of all subsets of $\bar{Z}$. This differs slightly – but *not significantly* – from the notion of a deterministic freedom measure introduced in Section 2.1 where a deterministic diversity was understood as a function whose arguments were subsets of Z rather than $\bar{Z}$. Note finally that the set of deterministic diversity measures is the set of *all* functions $\Lambda : 2^{\bar{Z}} \to \mathbb{R}_+$.

3 The Hybrid Measure

This section develops a hybrid measure that incorporates both the *intrinsic* and *instrumental* values of freedom. I first formalize the instrumental value of freedom in terms of preference for flexibility (Section 3.1). I then show that this notion of instrumental freedom is *formally* equivalent to the notion of a freedom measure that I developed in the previous section, and which I use to model intrinsic freedom (Section 3.2). However, I argue that instrumental and intrinsic freedom are not *substantively* equivalent (Section 3.3). I go on to argue that intrinsic freedom occupies an important place within prevailing value systems, and so we require a measure of freedom that integrates both intrinsic and instrumental values (Section 3.4). Section 3.5 develops such a hybrid measure.

3.1 Instrumental Value: Preference for Flexibility

A natural way to model the instrumental value of freedom is in terms of the flexibility that freedom allows an agent to adapt her decision to the contingencies and information that arise at the time of choice. Arrow (1995) proposed that freedom be conceptualized in this way. Others have developed the model of preference for flexibility without explicit reference to freedom (Kreps 1979, Dekel et al. 2001).

Let U be the set of possible utility functions on Z. Formally, $U = \mathbb{R}^Z$. For any $u = (u_z : z \in Z) \in U$, interpret u_z as the utility of outcome z. Let $p \in \Delta(U)$ be a probability measure over utility functions.

Suppose that at date 0, an agent must evaluate menus (of lotteries) M from which she will choose at date 1. At date 1, prior to her choice, her utility function in u will be realized, and she will choose optimally from M according to u. At date 0, the agent only has a probabilistic belief p over her date 1 utility functions. The agent evaluates a menu M via the expected utility it will generate from the perspective of date 0. Then, formally, the agent's date 0 evaluation of any menu of lotteries M is given by:

$$\iota_p(M) := \int_U \max_{\beta \in M} (u \cdot \beta) p(du),$$

where $u \cdot \beta = \sum_z u_z \beta_z$ is the agent's expected utility from lottery β when she has utility function u, and the expression $\max_{\beta \in M}(u \cdot \beta)$ encodes the assumption that at date 1 the agent will choose optimally given the utility function u that materializes (and she knows she will do so at date 0). $\iota_p(M)$ is the **indirect utility** of menu M. Call a probability measure $p \in \Delta(U)$ **admissible** if for all $M \in \mathcal{M}$, $-\infty < \iota_p(M) < \infty$. The phrase "preference for flexibility" refers to the fact that under this approach the agent will prefer larger menus because such menus allow her to adopt her choice to whichever utility function u materializes, as determined by the information and contingencies that turn out to be

relevant at the time of choice. Moreover, at date 0, the agent will typically prefer the flexibility embodied in menu M to the commitment at date 0 to any lottery β contained in M.

3.2 Formal Equivalence

Define $\widehat{U} := \{u \in U : u_z \geq 0, \forall z \in Z, \min_{z' \in Z} u_{z'} = 0\}$.

Proposition 2 *Let $\nu : \mathcal{M} \to \mathbb{R}$. There exists a stochastic freedom measure (λ, μ) such that $\nu = \nu_{\lambda,\mu}$ if and only if there exists an admissible $p \in \Delta(\widehat{U})$ such that $\nu = \iota_p$.*

Observe that for all $p \in \Delta(U)$, there exists $p' \in \Delta(\widehat{U})$ and a constant c such that $\iota_p = \iota_{p'} + c$, so the restriction to $p \in \Delta(\widehat{U})$ is not a substantive one.

Proposition 2 shows that the attribute and preference for flexibility approaches are formally equivalent, generalizing results about deterministic diversity/freedom measures (Nehring 1999, Nehring and Puppe 2002) to stochastic diversity/freedom measures. That is, the set of value functions generated by freedom measures is essentially the same as the set of value functions representing preferences of agents who value flexibility.

Corollary 1 *There exists a constant c such that $\nu + c$ is the value function generated by a stochastic freedom measure if and only if ν is continuous,*[27] *ν is monotone: $M \subseteq M' \Rightarrow \nu(M) \leq \nu(M'), \forall M, M' \in \mathcal{M}$, and:*

$$\nu\left(\alpha M + (1-\alpha)M'\right) = \alpha\nu(M) + (1-\alpha)\nu\left(M'\right), \quad \forall M, M' \in \mathcal{M}, \forall \alpha \in (0,1). \tag{5}$$

The proof of the corollary exploits the axiomatic characterization of preference for flexibility over menus of lotteries of Dekel et al. (2001). Formally, we define:

$$\alpha M + (1-\alpha)M' := \left\{\alpha\beta + (1-\alpha)\beta' : \beta \in M, \beta' \in M'\right\}.$$

Intuitively, for $\alpha \in (0,1)$, the menu $\alpha M + (1-\alpha)M'$ is the menu that effectively results in what I will refer to as the **random menu scenario**: With probability α, the agent faces menu M, and with probability $(1-\alpha)$, the agent faces menu M'; the set of lotteries in $\alpha M + (1-\alpha)M'$ is the set of lotteries that the agent could generate by some (pure) strategy for choosing out of M or M' – whichever materializes – in the random menu scenario.[28] Corollary 1 generalizes the characterization of the value functions corresponding

[27] Continuity is defined with respect to the Hausdorff topology on the set $\mathcal{M}$ of menus.

[28] It also follows from Proposition 2 and the analysis of Dekel et al. (2001) – this is fairly immediate from definitions even without the proposition – that if ν is the value function associated with a freedom measure, then $\nu(M) = \nu(\text{co}(M))$ where $\text{co}(M)$ is the convex hull of M; in other words no additional valuable freedom is generated by allowing the agent to randomize. So in the random menu scenario, the agent would not gain freedom if she were allowed to use randomized, rather and pure, strategies.

to deterministic diversity measures in Nehring (1999).[29] Intuitively (5) says that the value of freedom in the random menu scenario is equal to the expected value of the freedom of the menus generated in that scenario. (5) is essentially an independence condition analogous to that found in expected utility theory (see Dekel et al. (2001) for the corresponding axiom on preferences over menus).[30] The monotonicity condition that larger menus have (weakly) larger freedom value is immediate from definitions.

3.3 Substantive Inequivalence

While mathematically equivalent, the interpretations intended for the two approaches are not equivalent. The freedom measure approach is intended to capture the intrinsic value of having access to alternatives with certain attributes and the preference for flexibility approach is intended to capture the instrumental value a menu provides to adapt one's decision to factors relevant to the attainment of one's goals. Moreover the *information* one would need to construct value functions differs under the two approaches. In the preference for flexibility approach, one needs to know the likelihood that agents will have various preferences tomorrow (at the time of choice) and how agents trade off these possibilities

[29] Proposition 3 of Nehring (1999) establishes that ν is a value function associated with a deterministic diversity measure if and only if ν is monotone and totally submodular. Proposition 1 below shows how a stochastic diversity measure induces a deterministic diversity measure on deterministic menus. Corollary 2 establishes how Corollary 1 extends Proposition 3 of Nehring (1999) (Nehring (1999) defines diversity measures in such a way that if there exists a constant c such that $\nu + c$ is the value function corresponding to a diversity measure, then ν is also the value function of a diversity measure; the following corollary and its proof use the same terminological convention.)

Corollary 2 *A value function ν on the set of deterministic menus can be extended to the set of all menus so that it is continuous, monotone, and satisfies (5) if and only if on the set of deterministic menus, ν is monotone and totally submodular.*

Proof. First suppose that ν is value function on deterministic menus that can be extended to all menus so that it satisfies the appropriate properties. It is immediate that ν is monotone on deterministic menus. Then by Corollary 1 and Propositon 1 above, there exists $p \in \Delta(U)$ such that $\nu = \iota_p$. It then follows from Proposition 3 of Nehring (1999), ν is totally submodular on deterministic menus. Going in the other direction, suppose that ν is monotone and totally submodular on deterministic menus. Then by Proposition 3 of Nehring (1999), there exists $p \in \Delta(U)$ such that ν and ι_p coincide on the set of deterministic menus. But then by setting $\nu = \iota_p$ on all of $\mathcal{M}$, we get the desired extension. □

[30] Whereas I here derive (5) as a consequence of the preference for flexibility model (or alternatively, the stochastic freedom measure model), Dekel et al. (2001) start by justifying the independence axiom corresponding to (5), and then (employing some additional axioms) derive the preference for flexibility representation. Nevertheless their justification already appeals informally to the picture of an agent who evaluates today a menu from which she will choose tomorrow. For this justification, the reader is referred to Section 2.2 of their paper. A related justification could be provided from the perspective of stochastic freedom measures. The justification should have to steps (i) the freedom inherent in a menu should be independent of the procedure which generates the menu. So the agent has access to the menu because she will have the opportunity to select a strategy in the random menu scenario for some other reason is immaterial, and (ii) in comparing the random menu scenarios corresponding to $\alpha M + (1-\alpha)M'$ and $\alpha M + (1-\alpha)M''$, the relative freedom of the agent should be determined on the basis of the event that the agent will be confronted with different choices – that is the event that the agent will not face M. The argument seems equally plausible when applied to the hybrid measure to be presented in Section 3.5.

today (at time of menu evaluation). In contrast, the diversity approach requires judgments about the value of being able to choose alternatives with certain attributes.

The fact that two different sources of value can be expressed with the same mathematical formalism will help to provide a foundation for the hybrid measure in Section 3.5.

3.4 Shortcomings of the Instrumental Approach

At this point, the reader may wonder why we can't make do with the instrumental approach. This approach fits nicely within standard economics, but I argue that the instrumental approach fails to capture certain aspects of freedom deemed important by prevailing value systems.

One of the primary reasons for evaluating freedom is to decide who should be given which freedoms given that not everyone can control everything. To do so requires putting utilities on an interpersonally comparable scale. This will be done in Section 5.4. One interpretation of this scale is that it specifies who *cares more* about the various alternatives.[31] If the value of freedom is interpreted as instrumental along these lines, then it is difficult to capture common value judgments about who should have which freedom. Take, for example, the common belief that I should have control over my personal sphere (my body, my movements, my speech, ordinary activities I undertake in my home, etc.). Suppose that someone else – say Bob – happens to care more about the events occurring in my personal sphere than I do. This in itself would not justify transferring the right to control my personal sphere from me to Bob, at least under prevailing value systems.

To take an example of a different character, the freedoms associated with political speech should not be made contingent on the likelihood that that speech will be exercised. Suppose that the population can be split into two subpopulations, S and O, which are distinguishable on the basis of observable characteristics. Citizens in S *support* the government while those in O *oppose* the government. Citizens in S would never criticize the government. It would not be viewed as only a minor violation of freedom to pass a law that only citizens in O are allowed to criticize the government on the ground that citizens in S would never exercise this right (as opposed to the status quo that allows everyone to criticize the government). The instrumental approach to freedom would have a difficult time accounting for such judgments. Raz (1991) argues persuasively that it is difficult to justify the extent and strength of the protections of our freedoms of expression on the basis of their personal benefits to the possessors of these freedoms because most people exercise many aspects of such freedoms only in a limited way.

Generalizing the last example, sometimes an agent knows–and, moreover, everyone knows–how the agent would exercise her freedoms; there is no uncertainty about her future

[31] The interpretation of interpersonal comparisons in terms of who cares more is inessential. A similar argument could be constructed using only welfarism, the assumption that the social evaluation depends only on utilities or preferences.

behavior and no value for flexibility. Does this justify eliminating the agent's freedoms and imposing the outcome that she would choose? It seems not, at least according to prevailing value judgments. It can be as much a violation of freedom to restrict an agent who has decided on a course in such a way that she *must* choose that course as it would be to impose this course on an agent who has not decided it. We would not be indifferent between an arrangement in which everything is chosen for us by a benevolent planner who knows our preferences as well as we do and one in which we actually get to choose.

All of these examples show that, at least taken at face value, the instrumental approach fails to capture important intuitions about freedom. One important observation suggested by the above arguments is that *many fundamental rights do not appear to be particularly sensitive to the specific preferences of those who have them.*

One can attempt to argue that there is some implicit causal story and instrumental value judgement underlying all seemingly intrinsic evaluations of freedom.[32] One particular strategy would be to argue that the instrumental value does not accrue to the individual as option value as in the preference for flexibility story, but rather that a right to free speech is a *public good*.[33] (See Raz (1991) for advocacy of a public good account.) For example, when bad policies are being advocated or democracy is under threat, the ability of one person to speak out can protect all of society. Or the protection of free speech may contribute to the communication of novel ideas that help society flourish (Mill 1859). Such considerations are no doubt important for justifying freedoms. Such public good explanations can in general explain the insensitivity of at least some freedoms to the preferences of the individuals who possess them, since the good being protected accrues to society rather than to the particular individuals who possess them. Such public good explanations could provide an alternative foundation for a more narrowly "non-flexibility" as opposed to a "non-instrumental" component of freedom, which could be captured via a freedom measure.

The question remains whether public good explanations in conjunction with flexibility explanations and other instrumental explanations are exhaustive. Note first that such public good explanations seem to account better for some of the above examples than others. For example, the priority that an agent has over her own personal sphere even when someone else would care more to control it seems difficult to explain along these lines. Such would be a strained explanation to fit the theory. The public good explanation would also have a difficult time accounting for why we would prefer to choose rather than to have an omniscient

[32] Giving voice to a methodologically motivated version of this argument, Dasgupta (1995) writes, "It isn't enough to attack utilitarian theories merely by asserting that they glide over the claims of individual rights, and to produce hypothetical examples in which they move us in different ethical directions. Rights themselves need to be justified, and it would be remarkable if they could be justified in ways other than by an appeal to the human interests their recognition protects and promotes," writing elsewhere that "It is a deep and common intuition that freedom has intrinsic worth. But intuition is not always a reliable guide. What appears 'intrinsic' may have a deep instrumental root. As a research strategy, it makes sense to investigate whether freedom has instrumental worth."

[33] An account of the intrinsic value of freedom could also have a public good character.

benevolent planner choose everything on our behalf.[34]

There are further problems. Take the example with populations S and O. Here a public good explanation seems promising. Perhaps restricting the freedoms would set a dangerous precedent and make the population more vulnerable to government abuse in the future. Suppose however that there were a situation where, for whatever reason, the restriction would be unlikely to set a bad precedent – perhaps it is set by an unpopular government falling from power, or even assume that the law is likely to generate a backlash that would make such restrictions *less* likely in the future. It seems that in these circumstances the restriction would still be wrong.

This illustrates a general problem for instrumental explanations of intrinsic freedoms: The underlying causal stories can be be quite speculative and dependent on specific facts that may not always obtain when the freedom is valued, while the judgments that the corresponding freedoms should not be violated are often straightforward and insensitive to particular circumstances. Moreover, suppose that in some situation we devise some story about the bad consequences of restricting some freedom – say, restricting artistic speech. Perhaps it is argued that restricting artistic speech will lead to a restriction of political and other speech, causing the society to be less adaptive and innovative. Suppose, however, that the causal story turns out to be wrong. For example, other types of speech will not be affected, and the society will be just as innovative. Would we then conclude that it is acceptable to restrict artistic speech? If not, then the instrumental story would not be playing any role.[35] We may then go in search of another instrumental story and keep looking until we find one that seems to work, introducing an obvious bias. Even if we eventually find an instrumental story such that the causal connections we posit actually hold, these connections do not then justify the corresponding freedoms because our espousal of the freedoms apparently does not depend on the story.

One has to be careful not to take such arguments too far. We do not want to argue there can be no justification for intrinsic freedoms since we would hold on to the freedoms even if any purported justification for them fell apart. However, not all justifications are created equal. If we justify a freedom in terms of the intrinsic values such as the dignity and autonomy of the person, then we might be willing to agree that were it not for this dignity and autonomy, the freedom would not be important.

[34]Here one might invoke a learning story about the benefits of learning from the process of choosing, but the objection remains when the omniscient benevolent planner who will always choose for us already knows everything that we stand to learn.

[35]I am not arguing that consequentialist considerations could never justify restrictions of basic freedoms. Times of war may, for example, call for extreme measures. But when consequentialist considerations overrule a basic freedom, they must overcome a certain burden, which can be quite high. What accounts for this burden? I am arguing that we cannot explain this burden in purely consequentialist terms.

3.5 Foundation for a Hybrid Measure

The previous section argued that the instrumental account of freedom was not sufficient by itself. However, instrumental considerations are very important. This section provides a foundation for combining the instrumental and intrinsic values of freedom.

Fix an agent i and a set of outcomes Z. Suppose that we are already given both (i) the instrumental values of the agent, expressed as a probability measure $p \in \Delta(U)$ giving the distribution over i's utility function at the time of choice, and (ii) a freedom measure (λ, μ) giving the intrinsic value to i of having access to choice of alternatives with various attributes. p and (λ, μ) induce, respectively, the instrumental and intrinsic value functions on menus, $\nu^* = \iota_p$ and $\nu^\circ = \nu_{\lambda,\mu}$, as described in Sections 3.1 and 2.2.2. These value functions will generally be distinct as freedoms that are instrumentally valuable from the standpoint of i's preferences are generally not identical to freedoms judged to be intrinsically valuable for i (although the two may overlap).

The question before us now is how we integrate judgments about instrumental and intrinsic values into an overall judgment represented by a **hybrid** value function that unifies our judgments about the value of freedom. Proposition 3 grounds such a unification. Say that two value functions $\nu, \nu' : \mathcal{M} \to \mathbb{R}$ are **evaluation equivalent** if $\nu(M) \leq \nu(M') \Leftrightarrow \nu'(M) \leq \nu'(M'), \forall M, M' \in \mathcal{M}$. Evaluation equivalent value functions rank menus in the same way, or, in other words, represent the same preferences over menus.

Proposition 3 *Let the hybrid measure $\nu : \mathcal{M} \to \mathbb{R}$ satisfy (5) and:*

$$\nu^*(M) \leq \nu^*\left(M'\right) \text{ and } \nu^\circ(M) \leq \nu^\circ\left(M'\right) \Rightarrow \nu(M) \leq \nu\left(M'\right), \quad \forall M, M' \in \mathcal{M}, \tag{6}$$

with a strict inequality in the consequent whenever any of the inequalities in the antecedent are strict. Then there exists $\alpha \in (0,1)$ such that ν is evaluation equivalent to the value function ν' defined by:

$$\nu'(M) = \alpha\nu^*(M) + (1-\alpha)\nu^\circ(M), \quad \forall M \in \mathcal{M}. \tag{7}$$

ν' – like ν – satisfies (5).

To summarize, if the hybrid value function satisfies the independence condition (5), which, as shown by Corollary 1, is satisfied by both the intrinsic and instrumental value functions (see Section 3.2 for discussion), and the hybrid measure is positively responsive to instrumental and intrinsic freedoms (and to nothing else), then then the hybrid measure can be expressed as a weighted average of the instrumental and intrinsic values of freedom. For an alternative independent justification of (5), see footnote 30.

This result is a variant of Harsanyi's aggregation theorem (Harsanyi 1955), which was used to provide a foundation for utilitarianism. The substantive difference is in the in-

terpretation: Whereas Harsanyi takes the considerations to which the overall objective is responsive to be the preferences of individuals (which are instrumental values), I take these considerations to be the instrumental and intrinsic freedoms of a single individual. The result is formally similar to Theorem 6 of Kochov (2007), but the substantive interpretation is again quite different, as Kochov interprets the considerations being aggregated like Harsanyi as the preferences of different agents. Following Kochov, the proof is a corollary of the more general result of De Meyer and Mongin (1995) (see the Appendix).

4 Value of Life

This section applies the hybrid freedom measure ν to value of life calculations. Let $M = \{(r_1, w_1), \ldots, (r_n, w_n)\}$ be a menu of risk-wage pairs, where r_i is a risk, or mortality probability, and w_i is a wage. Suppose the attributes assumed to have intrinsic value are those associated with survival. Specifically, let $A(q)$ be the attribute that the agent survives with probability at least q, and assume that for all q, the intrinsic value associated with $A(q)$ is 1. Recall from Section 2.2.2 that this is interpreted as the marginal value associated with increasing the probability q that the agent survives. One can spell out the details of the freedom measure such that:[36]

$$\nu^\circ(M) = \max_{(r,w)\in M}(1-r),$$

or in other words, the intrinsic value of any menu of risk-wage pairs is the maximum survival probability allowed by some alternative in the menu.

The reader may or may not share the view that survival is intrinsically valuable, and moreover is the unique intrinsically valuable property in this setting. The purpose of this section is not to take a specific stance on how the value of life should be assessed, but rather to illustrate how evaluation can be done when the value of intrinsic freedom is taken seriously.

For simplicity, let us assume that there is no ex ante uncertainty about the agent's preference over risk-wage pairs. Let $u(r, w)$ be the agent's expected utility of (r, w). Then

[36]Formally, let W be a finite collection of wages. Then we can define $Z = \{0, 1\} \times W$, where 1 represents the situation in which the agent survives and 0 represents the situation where the agent dies. Then a risk-wage pair (r, w) is formally a lottery that puts probability r on $(0, w)$ and probability $1-r$ on $(1, w)$. $u(r, w)$ is the agent's expected utility from this lottery. Define $v = (v_z : z \in Z) \in V$ to be the vector such that $v_z = x$ if z is of the form $(1, w)$ and $v_z = -x$ if z is of the form $(0, w)$, where $x > 0$ is a real number chosen so that $\|v\| = 1$. Then $A(q)$ defined in the text can more formally be written as $A(v, q)$ as in section 2.2.1. Define the reference measure μ so that it puts probability 1 on v, and define λ so that $\lambda(v) = 1$. Then for menus M of risk-wage pairs:

$$\nu^\circ(M) = \int_{\mathcal{A}(M)} \tilde{\lambda} d\tilde{\mu} = \int_{[0,1]} 1[A(q) \cap M \neq \emptyset] dq = \max_{(r,w)\in M}(1-r),$$

where $1[\cdot]$ is the indicator function.

the instrumental value of a menu is given by the indirect utility:

$$\nu^*(M) = \max_{(r,w)\in M} u(r,w)$$

Then the hybrid value is given by:

$$\nu(M) = \alpha \max_{(r,w)\in M} u(r,w) + (1-\alpha) \max_{(r,w)\in M} (1-r),$$

where $\alpha \in (0,1)$. To make the example a bit more specific, let w^h, w^m, w^ℓ be respectively, high, medium, and low wages, so that: $w^h > w^m > w^\ell$. Assume that the agent's utility function u is such that:

$$u\left(\frac{1}{10}, w^h\right) > u\left(\frac{1}{10}, w^m\right) = u\left(0, w^\ell\right)$$

The agent obviously prefers a high risk-high wage job to a high risk-medium wage job. The agent is also indifferent between a high risk-medium wage job and a low risk-low wage job. Table 1 provides three menus along with their hybrid values: M_1 contains both a

	M	$\nu(M)$
$M_1:$	$\left\{\left(\frac{1}{10}, w^h\right), \left(0, w^\ell\right)\right\}$	$\alpha u\left(\frac{1}{10}, w^h\right) + (1-\alpha)$
$M_2:$	$\left\{\left(\frac{1}{10}, w^h\right)\right\}$	$\alpha u\left(\frac{1}{10}, w^h\right) + (1-\alpha)\frac{9}{10}$
$M_3:$	$\left\{\left(0, w^\ell\right)\right\}$	$\alpha u\left(0, w^\ell\right) + (1-\alpha)$

Table 1: Menus of Risks and Wages

high risk-high wage job and a low risk-low wage job, M_2 contains only the former, and M_3 contains only the latter. Assume that the agent evaluates menus only according to their instrumental values and not according to the intrinsic freedom they provide. (See Section 7.1 for a discussion of the possibility that rather than only being viewed as socially valuable, agents themselves partially or fully internalize the value of intrinsic freedoms). Then the agent is indifferent between M_1 and M_2, since they both contain her favorite combination $\left(\frac{1}{10}, w^h\right)$. The agent prefers both of these menus to M_3 which does not contain this combination. In contrast, the hybrid measure is not indifferent, but prefers M_1 to M_2. This is because M_1 allows more freedom. It is also possible, depending on the value of α and other parameters that the hybrid measure prefers M_3 to M_2, because of the additional freedom it grants, making an evaluation contrary to the agent's preference.

To *impose* a mortality risk of $\frac{1}{10}$ – that is to give the agent no choice but to face this risk – given status quo $\left(0, w^\ell\right)$, we must compensate the agent by more than $w^m - w^\ell$, the additional wage that would make the agent indifferent, in order to make the hybrid measure indifferent between the new situation and the status quo. This is because by imposing the risk, we not only harm the agent, we rob her of her liberty. It is intuitive that to impose a

risk on an agent, we must compensate her by more than she demands. This is because if she were to freely make the choice to take the risk, she would be responsible for the risk, and would not deserve additional compensation.

Judgments similar to the above may apply to cost-benefit analysis in other domains as well. In principle, taking the intrinsic value of freedom seriously could have a drastic effect on the policies recommended by cost-benefit analysis.[37] The hybrid measure evaluates not only welfare but the *procedural* aspects of mechanisms: namely whether the agent chooses the risk, or whether it is imposed on her. It is important to note, however, that what counts as an "imposition" depends on which freedoms are viewed as intrinsically valuable. In this example, we assumed that survival is intrinsically valuable, but receiving a high wage is not. Had we assumed otherwise, we might have obtained different results. For further discussion of this example, see Section 7.3.

5 Evaluating Freedom in Interactive Settings

I now extend the model to a situation where the set possible outcomes Z is not under the control of a single agent, but is rather jointly controlled by a collection I of agents. As above, each agent may have one of many preferences over alternatives. Formally, for each agent i, there is now a distinct probability measure p_i over the set of possible utility functions U. The different probability measures p_i are independent, and there is also common knowledge of the p_i.

To illustrate, consider the communal bicycle example (CB) from the introduction, and suppose that the two agents are Ann and Bob. $\mathcal{L}$ is a set of locations. An outcome in Z specifies which agent (if any) takes the bicycle, and if someone takes the bicycle, to which location L in $\mathcal{L}$ the agent travels. Each agent wakes up with a desire to travel to some location, or to stay home, and this is captured by the utility function u_i that materializes. The probability (from the standpoint of the night before) that the agent will have various desires for travel is captured by the probability measures p_i. In this particular example, the assumption that each agent cares only about her own travel can be captured by assuming that with probability 1, p_i selects a utility function that does not depend on the other agent's location.

5.1 Social Interactions

To model social interactions, we need to specify both the *rules of the game* that govern interaction and the *behavior that prevails* given the rules.

[37] For related arguments see Zamir and Medina (2008), Zamir and Medina (2010), and Lowry and Peterson (2011).

5.1.1 Rules of the Game

I model the rules of the game via a **game form** $\Gamma = ((S_i : i \in I), g)$, where S_i is a finite collection of actions for player i and $g : \times_{i \in I} S_i \to \Delta(Z)$ is an **outcome function** that maps an action profile (that is, a list of actions, one for each agent) to a lottery over outcomes.

In CB, the set of actions corresponds to the set of locations, as well as the option of choosing not to use the bicycle. The action corresponding to location L corresponds to attempting to go to location L. If only one agent attempts to use the bicycle to go to location L and the other opts not to use the bicycle, the outcome function g specifies that the agent who attempts to go to L goes to L, and the other stays home. If one agent attempts to go to location L and the other to L', one could assume that each one will get her/his way with probability $\frac{1}{2}$ and the other will stay home. For simplicity, I will henceforth assume that if both attempt to use the bicycle, then Bob will get there first and will get his way, and Ann will stay home.

5.1.2 Prevailing Behavior

I take the stance that to assess freedom, it is not sufficient to just know the formal rules that govern the social interaction, but it is also necessary to know what behavior prevails. Aside from its intuitive plausibility, this stance is necessary to extend the approach of Sections 2-3 from individual decisions to social interactions.

Prevailing behavior is modeled via the notion of a strategy: A **strategy** for agent i is a measurable function $\sigma_i : U \to \Delta(S_i)$.[38] That is, for each utility function u_i that could materialize for i, σ_i specifies what action i will choose conditional on u_i, and σ_i allows that i could randomize over actions. Notationally, $[\sigma_i(u_i)](s_i)$ is the probability that agent i will select action s_i when she has utility function u_i. Built into the framework is the assumption that at the time that i chooses her action she knows her own utility function in U, but she does not know the utility functions of the other agents; so σ_i does not allow i to base her decision on the utility functions of others. So far, the general approach is standard from the framework of Bayesian games. As a matter of notation, σ denotes a strategy profile (a list of strategies, one for each agent). Let σ_{-i} be a strategy profile for all agents other than i. So σ_{-i} (together with the profile of probability measures p_{-i} other than p_i) determines the distribution of behavior that i faces. Using standard notation, I write $\sigma = (\sigma_i, \sigma_{-i})$.

In CB, σ_i specifies agent i's choice of a location to attempt to visit or a decision to stay home as function of the preferences u_i she discovers when she awakens in the morning.

[38] For our purposes, it would be sufficient if σ_i had as its domain only the support of p_i.

5.2 Effective Menus

The key idea is that the rules of the game combined with prevailing behavior determine an *effective menu* for each agent; in other words, what an agent can effectively choose depends on the rules and the behavior of others. Applying the theory of Sections 2-3 to agents' effective menus allows me to extend the analysis from decisions to interactions. I now spell out the details.

Define $\beta_i(s_i, \sigma_{-i})$ to be the lottery that would result if agent i selected the (pure) action s_i and all other players used strategy σ_{-i}. For example, abbreviate Ann and Bob by a and b respectively, and suppose that in CB, Bob's strategy σ_b is such that he attempts to use the bicycle $\frac{2}{3}$ of the time, and recall that if Both attempt to use the bicycle, Bob succeeds. Then if $s_a = L$ is Ann's action of attempting to go to L, $\beta_a(L, \sigma_{-a})$ is the lottery according to which Ann goes to L with probability $\frac{1}{3}$ and stays home with probability $\frac{2}{3}$.[39] This is so, because Ann will only succeed in going to L $\frac{1}{3}$ of the time.

We would like to interpret $\beta_i(s_i, \sigma_{-i})$ as a lottery that is available to the agent at the interim stage – that is, the time of choice – after i has learned her own preferences but before she has learned those of the other players.[40] Because preferences are independent, $\beta_i(s_i, \sigma_{-i})$ does not depend on which preferences u_i she learns she has, and moreover, $\beta_i(s_i, \sigma_{-i})$ is also the lottery that would result if i committed ex ante before i learns her preferences to choosing s_i while all other players play according to σ_{-i}. Formally, $\beta_i(s_i, \sigma_{-i})$ is the lottery that assigns to outcome z probability

$$\int_{U^{I\setminus i}} \sum_{\left\{s_{-i}=(s_j)_{j\in I\setminus i}\in S_{-i}\right\}} \prod_{j\in I\setminus i} [\sigma_i(u_j)](s_j)[g(s_i, s_{-i})](z)p_{-i}(du_{-i}),$$

where $u_{-i} \in U^{I\setminus i}$ is a generic profile of utility functions for all agents other than i, $p_{-i} := \times_{j\in I\setminus i} p_j, S_{-i} := \times_{j\in I\setminus i} S_j$, and $[g(s_i, s_{-i})](z)$ is the probability assigned to z by outcome function g when i selects action s_i and other agents select actions s_{-i}.

A key concept in the analysis is agent i's **effective menu**, defined as $M_i(\Gamma, \sigma_{-i}) = \{\beta_i(s_i, \sigma_{-i}) : s_i \in S_i\}$.[41] That is, $M_i(\Gamma, \sigma_{-i})$ is the set of lotteries from which i can effectively choose by varying her choice of actions, holding fixed the behavior of others.[42] I illustrate with CB: If Bob's strategy is such that he ultimately takes the bicycle with probability π,[43] then if Ann attempts to take the bicycle, she will be successful with probability

[39] Strictly speaking, $\beta_a(L, \sigma_{-a})$ also specifies the distribution of places Bob will go when he takes the bicycle, but this can be ignored for the purpose of the discussion.

[40] If the agent ever learns the preferences of the other agents, it is only after she has chosen an action.

[41] $M_i(\Gamma, \sigma_{-i})$ is also a function of p_{-i}, although I have suppressed this argument.

[42] One might think that the effective menu should be the menu that results from letting i randomize over actions rather than merely selecting pure actions. However, one can show that the hybrid measure is such that the value of any menu M is equal to the value of the convex hull of M; so this alternative approach to defining the effective menu would not lead to a difference in evaluation.

[43] Under this strategy, whether Bob takes the bicycle may depend on his utility function u_b; the total

$(1-\pi)$. For each location L, $M_a(\Gamma, \sigma_{-a})$ contains the lottery according to which Ann travels to L with probability $(1-\pi)$, and remains home with probability π.[44] ($M_a(\Gamma, \sigma_{-a})$ also contains the lottery where Ann stays home with probability 1 because Ann has the option not to take the bicycle). As the lotteries effectively available to Ann depend on π, this example illustrates how Ann's effective menu depends on Bob's behavior.

5.3 Instrumental Freedoms in an Interactive Setting

This section defines instrumental freedoms in an interactive setting, and discusses some of the attendant subtleties.

I propose to measure the instrumental freedom of an agent in a social interaction as the instrumental value of the agent's effective menu in that social interaction. In CB, if Bob uses the bicycle with probability π, then the value of Ann's instrumental freedom is the value that Ann would achieve were she to choose from the menu that allowed her to go to any location with probability $(1-\pi)$ (and stay home with the remaining probability).[45] Formally, the instrumental value of Ann's freedom in a social interaction is $\hat{\nu}_i^*(\Gamma, \sigma)$, defined by:

$$\hat{\nu}_i^*(\Gamma, \sigma) = \iota_{p_i}\left(M_i\left(\Gamma, \sigma_{-i}\right)\right). \tag{8}$$

The above definition makes an agent's instrumental freedoms depend on the prevailing behavior of others – the behavior the agent can reasonably expect – insofar as this behavior determines the menu effectively facing the agent. This is as it should be. My freedom to walk the streets safely depends not only on the laws but also on the behavior that prevails in consequence. If others threaten me, I am not so free. In principle one could instead take a *purely* procedural approach according to which only the formal rules governing interaction – and not actual behavior – determine agent's freedoms. I do not think it is possible to reasonably assess freedom in abstraction from how the behavior of others impinges on one's choices.

The above approach makes the agent's ability to evaluate her own freedoms dependent on the formation of correct expectations about the distribution of behaviors by others. Individuals might not always form such correct expectations, but this is a reasonable modeling assumption, especially in stable social environments. The correct predictions alluded to above have an equilibrium flavor, and indeed we shall see shortly that equilibrium plays an important conceptual role in assessing freedom in an interactive setting.

To understand the role of equilibrium in the analysis, it is necessary to introduce some definitions. Let $\widehat{U}_i\left(s_i, \sigma_{-i} | u_i\right)$ be i's (interim) expected utility conditional on learning that

resulting probability that he takes the bicycle is π.

[44] Again, strictly speaking, these lotteries also contain information about Bob's location: See footnote 39.

[45] The menu also contains the option of staying home with probability 1.

her utility function is u_i if i selects action s_i and other agents play according to strategies σ_{-i}. Let $U_i(\sigma)$ be i's ex ante expected utility – that is, i's expected utility before she learns her utility function u_i – under the strategy profile σ. Formally:

$$\widehat{U}_i(s_i, \sigma_{-i}|u_i) = \int_{U^{I\setminus i}} \sum_{\left\{s_{-i}=(s_j)_{j\in I\setminus i}\in S_{-i}\right\}} \prod_{j\in I\setminus i} [\sigma_j(u_j)](s_j)\,[g(s_i, s_{-i})](z)\,u_i(z)\,p_{-i}(du_{-i}),$$

$$U_i(\sigma) = \int_U \sum_{s_i\in S_i} [\sigma_i(u_i)](s_i)\,\widehat{U}_i(s_i, \sigma_{-i}|u_i)\,p_i(du_i),$$

where $u_i(z)$ is the utility that utility function u_i assigns to outcome z.

An equilibrium is a situation in which every agent's behavior is optimally adapted to the behavior of all other agents. It can be reasonable to expect equilibrium to obtain in stable environments where each agent has the opportunity to learn the distribution of behavior of others and adapt her behavior accordingly. Formally, a **Bayesian Nash equilibrium (BNE)** is a strategy profile $\sigma^* = (\sigma_i^* : i \in I)$ such that:[46]

$$\forall i \in I, \forall u_i \in U, \forall s_i \in S_i, [\sigma_i^*(u_i)](s_i) > 0 \Rightarrow s_i \in \arg\max_{s_i'\in S_i} \widehat{U}_i\left(s_i', \sigma_{-i}|u_i\right)$$

This means that each agent's expected utility maximizes her expected utility given others' strategies. If σ^* is a BNE, then we refer to $M_i(\Gamma, \sigma_{-i}^*)$ as i's **equilibrium menu**.

Say that under σ, agent i **realizes the instrumental value of her effective menu** if $U_i(\sigma) = \hat{\nu}_i^*(\Gamma, \sigma)$. That is to say, i's choice according to σ_i is actually such that i's expected utility is equal to the instrumental value of the social interaction. The following observation provides a fundamental relationship between the realization of instrumental values and equilibrium.

Observation 1 *All agents realize the instrumental values of their effective menus in σ^* if and only if σ^* is a BNE.*

If σ is not a BNE, then there must be some agent i such that $\hat{\nu}_i^*(\Gamma, \sigma) \neq U_i(\sigma)$. Even if σ is not a BNE, it is still coherent to interpret $\hat{\nu}_i^*(\Gamma, \sigma)$ as the (maximum) value i has the *potential* to achieve in the interaction – holding fixed others' behavior. Because I do henceforth restrict attention to BNE, $\hat{\nu}_i^*(\Gamma, \sigma)$ is equal to i's *actual* expected utility in the interaction. The interesting question of how to adapt the analysis to interactions that are out of equilibrium is left to future work.

Formally, define a **social arrangement** as a pair (Γ, σ) where Γ is a game form and σ is a BNE of Γ. Whether σ is a BNE of Γ depends on the profile of probability measures $p = (p_i : i \in I)$ over utility functions. Let $\mathcal{SA}_p$ be the set of social arrangements given p.

[46] It would be sufficient to impose the following condition only on u_i in the support of p_i rather than on all u_i in U.

For any pair of social arrangements (Γ, σ) and (Γ', σ') in $\mathcal{SA}_p$ and number $\alpha \in (0,1)$, it is convenient to define the **mixture** $\alpha(\Gamma, \sigma) + (1-\alpha)(\Gamma', \sigma')$ to be the social arrangement in which agents play Γ with probability α and Γ' with probability $1-\alpha$, – all are informed of whether Γ or Γ' is to be played – and the agents use strategy profile σ conditional on playing Γ, and strategy profile σ' conditional on playing Γ'. A more formal definition is given in the proof of Proposition 4 in the Appendix. It is straightforward to establish that if (Γ, σ) and (Γ', σ') are social arrangements, then $\alpha(\Gamma, \sigma) + (1-\alpha)(\Gamma', \sigma')$ is also a social arrangement (i.e., an element of $\mathcal{SA}_p$).

Throughout the preceding parts of Section 5, I have written as though the value of intrinsic freedom is not internalized in behavior. In the next subsection, however, I will use the value of intrinsic freedom (along with the instrumental value) to *evaluate* social interactions. This is completely analogous to the single agent case presented in Section 3 as I have interpreted it, although other interpretations may be compatible with the formalism. This assumption about internalization, as well as the question of whether it is necessary, is discussed in depth in Section 7.1.

5.4 Evaluating Social Interactions

This section develops a hybrid measure for social interactions, incorporating both instrumental and intrinsic values, extending the approach of Section 3.5 from menus to social interactions.

To construct a hybrid measure, it is necessary first to extend the intrinsic value functions of Section 2.2 to social interactions. For each agent i, I posit a (stochastic) freedom measure (λ_i, μ_i) encoding the intrinsic values of i's intrinsic freedoms. Each freedom measure induces a menu value function ν_i° via (3). I extend intrinsic values to social interactions in the same way that I did for instrumental values (see (8)): The intrinsic value of i's freedom in a social interaction $\hat{\nu}_i^\circ(\Gamma, \sigma)$ is defined as the intrinsic value of i's effective menu in that interaction. Formally, $\hat{\nu}_i^\circ(\Gamma, \sigma) = \nu_i^\circ\left(M_i\left(\Gamma, \sigma_{-i}\right)\right)$. It is also convenient to define the function $c_i(\Gamma, \sigma) = \nu_i^\circ(\bar{Z}) - \hat{\nu}_i^\circ(\Gamma, \sigma)$. Recall that $\bar{Z}$ is the set of deterministic lotteries that allow complete control of the outcome. So $c_i(\Gamma, \sigma)$ is the difference between the maximum possible intrinsic value that the agent could achieve if she were to have complete control of the outcome[47] and the value that she obtains in (Γ, σ). It is generally not possible for everyone to control everything; so $c_i(\Gamma, \sigma)$ represents the intrinsic costs associated with the sacrifice of some of i's control in (Γ, σ).

Suppose that we are given the probability measures $p = (p_i : i \in I)$ and the freedom measures $((\lambda_i, \mu_i) : i \in I)$ inducing the value functions $\hat{\nu}_i^*$ and $\hat{\nu}_i^\circ$ on $\mathcal{SA}_p$. Proposition 4 – the analogue for interactions of Proposition 3 for menus – provides a foundation for a hybrid

[47] It follows from the analysis of Section 2 that the agent achieves just as much intrinsic value with $\bar{Z}$ as she would with the menu $\Delta(Z)$ containing all lotteries.

measure on the set of social interactions. Say that two value functions $\hat{\nu}, \hat{\nu}' : \mathcal{SA}_p :\to \mathbb{R}$ are **evaluation equivalent** if $\hat{\nu}(\Gamma, \sigma) \leq \hat{\nu}(\Gamma', \sigma') \Leftrightarrow \hat{\nu}'(\Gamma, \sigma) \leq \hat{\nu}'(\Gamma', \sigma')$ for all $(\Gamma, \sigma), (\Gamma', \sigma') \in \mathcal{SA}_p$. This is analogous to the definition of evaluation equivalence for menu value functions in Section 3.5.

Proposition 4 *Suppose that the social objective $\hat{\nu} : \mathcal{SA}_p \to \mathbb{R}$ satisfies:*

$$\hat{\nu}(\alpha(\Gamma, \sigma) + (1-\alpha)(\Gamma', \sigma')) = \alpha\hat{\nu}(\Gamma, \sigma) + (1-\alpha)\hat{\nu}(\Gamma', \sigma'), \quad \forall (\Gamma, \sigma), (\Gamma', \sigma') \in \mathcal{SA}_p \tag{9}$$

If $\hat{\nu}$ depends positively and only on the intrinsic and instrumental freedoms of each of the agents:

$$\left(\hat{\nu}_i^*(\Gamma, \sigma) \leq \hat{\nu}_i^*\left(\Gamma', \sigma'\right) \text{ and } \hat{\nu}_i^\circ(\Gamma, \sigma) \leq \hat{\nu}_i^\circ\left(\Gamma', \sigma'\right), \forall i \in I\right) \Rightarrow \hat{\nu}(\Gamma, \sigma) \leq \hat{\nu}\left(\Gamma', \sigma'\right), \\ \forall (\Gamma, \sigma), \left(\Gamma', \sigma'\right) \in \mathcal{SA}_p, \tag{10}$$

with a strict inequality in the consequent whenever any of the inequalities in the antecedent are strict. Then there exist positive real numbers w_i^*, w_i° (for $i \in I$) and κ such that $\hat{\nu}$ is evaluation equivalent to $\hat{\nu}'$ defined by:

$$\hat{\nu}'(\Gamma, \sigma) = \underbrace{\sum_{i \in I} w_i^* \nu_i^*(\Gamma, \sigma)}_{\text{utilitarian welfare}} - \underbrace{\kappa \sum_{i \in I} w_i^\circ c_i(\Gamma, \sigma)}_{\text{costs of rights violations}}. \tag{11}$$

$\hat{\nu}'$ – like $\hat{\nu}$ – satisfies (9).

The proposition provides a foundation for the natural and intuitive idea that the value of freedom in a social interaction consists in the utilitarian sum of instrumental values of freedoms realized in that interaction minus the sum of costs associated with rights violations – or curtailments of intrinsically valuable freedoms. Thus, consequentialist considerations are traded off against considerations of a more deontic character.[48] In (11), the weights w_i^* control the relative importance of the instrumental values, and the weights w_i° control the relative importance of the freedom violations of different agents. κ is a weight that can be used to control the relative importance of instrumental values as opposed to freedom violations.[49]

I now discuss the assumptions under which the proposition holds. (10) is the analogue of (6), and says that the social evaluation is increasing in both the instrumental and intrinsic freedoms of individual agents. (9) is an independence condition for social interactions

[48] Relatedly, Zamir and Medina (2008) propose incorporating deontological constraints into cost-benefit analysis.

[49] It would of course be possible to eliminate κ and absorb the weight that would otherwise be carried by κ uniformly into the weights w_i°. I keep the variable κ because for comparative statics, it can be useful to have a variable that shifts weight toward or away from intrinsic freedoms generally.

analogous to the condition (5) for menus. The justification is also analogous and may depend on the fact that each ν_i^* and ν_i° individually satisfies the condition in (9) and we may want to impose the same condition on the overall value judgment, or on an argument analogous to that in footnote 30 (see also Section 3.5). In a multi-agent context, however, additional considerations may be relevant for its assessment, specifically, considerations of fairness: While intuitive, (9) might be inconsistent with some judgments on fairness.[50] Note however that the current purpose is to *expand* the set of normative considerations relative to purely utilitarian evaluation. Here I explicitly incorporate the intrinsic value of freedom; future work could combine this with considerations of fairness. The current procedure maximizes the analogy between the single and multi-agent cases as considerations of fairness do not arise in the single-agent case. Incorporating fairness would bring an additional ingredient into the mix but would not change the basic message that instrumental and intrinsic freedoms must be traded off.

As with Proposition 3, the proof (in the Appendix) resembles that of Harsanyi's aggregation theorem. What is interesting here is that once one takes intrinsic freedom seriously, the same sort of formal arguments that Harsanyi used to justify utilitarianism justify trading off utilities against freedom in the form (11).

I note that there is an alternative derivation of the social value function (11). We could have started by deriving a hybrid value function ν_i for each agent via Proposition 3, extended this to a measure on social arrangements $\hat{\nu}_i$ (as we did for ν_i^* and ν_i°), and then, instead of imposing the monotonicity condition (10) on the value functions $\hat{\nu}_i^*$ and $\hat{\nu}_i^\circ$ (for $i \in I$), we could have imposed an analogous monotonicity condition on the $\hat{\nu}_i$'s. This would have equivalently led to a social objective of the form (11). This version of the procedure separates the value judgements involved in aggregating intrinsic and instrumental freedoms within an individual from the value judgments aggregating overall freedoms across individuals.

5.4.1 A Limiting Case: When Only Instrumental Values Matter

Consider the limiting case in which intrinsic values are **empty**: that is, $\nu^\circ(M) = 0$ for all $M \in \mathcal{M}$. Then the social objective of Proposition 4 is purely instrumental and we have:

Observation 2 *Assume that intrinsic values are empty. Then in the social aggrangement $(\Gamma, \sigma) \in \mathcal{SA}_p$, the value of freedom according to the social objective of Proposition 4 is a weighted sum of agents' ex ante expected utilities in the BNE σ.*

[50] Suppose that there are two agents, 1 and 2. Let (Γ^i, σ^i) be the social arangement in which agent i can select any outcome in Z – and always chooses optimally according to her realized preferences u_i – and in which the other agent has no power. Suppose that we are indifferent between (Γ^1, σ^1) and (Γ^2, σ^2). Then (9) implies that we will be indifferent between (Γ^1, σ^1) and the mixture $\frac{1}{2}(\Gamma^1, \sigma^1) + \frac{1}{2}(\Gamma^2, \sigma^2)$, in which each agent is given the opportunity to choose with equal probability. However, it would be reasonable to prefer the latter on the ground that it is more fair.

5.5 Partial rankings

Finally a word is in order about the completeness of the objective (11). It is likely that in reality, we will only be able to make crude and incomplete judgments about the values of various freedoms. So what we will need is a partial ranking of social arrangements rather than a total ranking as in (11). One formal alternative to represent this would be to take the consensus of a set of objectives of the form (11) in which the costs of violating freedoms varies within some range. I do not pursue this alternative here, but only in the interest of simplicity of exposition.[51]

6 Trade of Control in the Personal Sphere

Two decisions must allotted among Ann and Bob: *What color shirt should Ann wear?* and *What color shirt should Bob wear?* This is of course highly stylized. We do not ordinarily debate who should control whose shirt color. The shirt colors are to be understood as allegorical, representing the personal spheres of the agents. The example is chosen specifically to illustrate the potential conflict between instrumental values and the priorities that we often feel agents should have over personal decisions. A similar example was introduced and studied by Gibbard (1974).

Instrumental Values. Ann and Bob are abbreviated a and b respectively. When i is either Ann or Bob, $-i$ is the other agent. An **outcome** is a pair (z_a, z_b). z_a is the color – either *green* or *red* (abbreviated as g and r respectively) – of the shirt that Ann wears, and z_b is the color – g or r –of the shirt that Bob wears. Z is the set of possible outcomes. Each agent may be of four possible **types**, depending on which shirt color that agent prefers that she wears and which shirt color she prefers that the other agent wears. So if z_i' and z_{-i}' are shirt colors, an agent i of type $\bar{z}' = (z_i', z_{-i}')$ prefers that she wear shirt color z_i' and the other agent wear shirt color z_{-i}'. If z and z' are two shirt colors, then let $m(z, z')$ be a function that specifies whether they match: $m(z, z') = 1$ if $z = z'$ and $m(z, z') = -1$ if $z \neq z'$. Let γ_i be a positive number representing how much i cares about the other's shirt. If i is of type (z_i', z_{-i}') and i wears shirt color z_i while $-i$ wears shirt color z_{-i}', i's utility is:

$$u_i(z_i, z_{-i}; \gamma_i, z_i', z_{-i}') = m(z_i, z_i') + \gamma_i m(z_{-i}, z_{-i}').$$

That is, i receives (1) a utility of 1 if she wears the shirt color that she prefers and -1 if she wears the shirt color she prefers not to wear, and (2) an additional utility of γ_i if the other agent wears the shirt that she prefers the other agent to wear and $-\gamma_i$ if the other agent wears the shirt she disprefers the other agent to wear.

[51] Sher (2014) explores one approach to establishing incomplete comparisons, when value judgments concern what properties of outcomes are significant but are neutral about which specifications of those properties are better.

We would like to evaluate alternative arrangements from the perspective of a time before each agent knows which shirt colors she would like the two agents to wear. Each of the four possible types (z_i', z_{-i}') that i could be has probability 1/4. The probabilities of the types of the two agents are independent. Following the program set forth in previous sections, to evaluate arrangements, we must evaluate menus of lotteries over possible outcomes. For a menu M of lotteries over the shirt color choices of the two agents, the instrumental value that agent i attains from menu M is: (where we write $\beta(\bar{z})$ instead of $\beta_{\bar{z}}$ – see Section 2.2)

$$\nu_i^*(M) = \frac{1}{4} \sum_{\bar{z}' \in Z} \max_{\beta \in M} \sum_{\bar{z} \in Z} u_i(\bar{z}; \gamma_i, \bar{z}') \beta(\bar{z}). \tag{12}$$

Cases. We focus on two cases:

- **Bob intrusive (BobInt).** Bob prefers to control the color of Ann's shirt than to control the color of his own shirt. Ann prefers to control the color of her own shirt than to control the color of Bob's shirt. Formally, $0 < \gamma_a < 1$ and $\gamma_b > 1$.
- **Both intrusive (BothInt).** Both agents prefer to control the color of the other agent's shirt than to control their own. Formally, $\gamma_a > 1$ and $\gamma_b > 1$.

Intrinsic Values. We consider two possible specifications of intrinsic values: *empty priorities* and *personal sphere priorities*, corresponding to superscripts $\emptyset$ and ps respectively. Under **empty priorities**, the freedom measures for the two agents are uniformly zero: $\lambda_i^{\emptyset} \equiv 0$. In this case, only instrumental values matter.

Under **personal priorities**, each agent has priority over her own personal sphere. The attributes that are relevant to i in this case are those of the form: *The probability that i wears shirt color z is at least q.* Let us denote this attribute as $E_i(z, q)$, where $q \in (0, 1)$. Translating this into the notation of Section 2.2.1, we can write $E_i(z, q) = A\left(v^{i,z}, q\right)$, where $v^{i,z}$ is a vector of the form $v^{i,z} = \left(v^{i,z}_{(z_i', z_{-i}')} : (z_i', z_{-i}') \in Z\right)$ such that $v^{i,z}_{(z_i', z_{-i})} = 1/2$ if $z = z_i'$ and $v^{i,z}_{(z_i', z_{-i}')} = -1/2$ if $z \neq z_i'$. The reference measure μ_i^{ps} is the "uniform" discrete measure that puts an equal weight of 1 on each of the two vectors $v^{i,g}$ and $v^{i,r}$, and puts a weight of 0 on all other vectors in V. The freedom measure is then defined so that $\lambda_i^{ps}\left(v^{i,z}\right) = 1$ for $z = g, r$ and $\lambda_i^{ps}(v) = 0$ for all other v. In sum, attributes of the form $E_i(z, q)$ are given a weight of 1, and all other attributes are given a weight of 0. The induced intrinsic value function is then given by:

$$\nu_i^{\circ,ps}(M) = \sum_{z_i \in \{g,r\}} \max_{\beta \in M} \beta_i(z_i) \tag{13}$$

where for any lottery β and any shirt color z_i, $\beta_i(z_i) = \sum_{z_{-i} \in \{g,r\}} \beta(z_i, z_{-i})$ is the probability that lottery β assigns to i getting a z_i color shirt. Thus, the intrinsic value of a menu for i is the sum of the maximal achievable probabilities of each shirt color for i within that menu,

ranging from a minimum possible value of 1 for a menu containing only a single lottery to a maximum possible value of 2 for a menu that allows i to perfectly control her shirt color.

Mechanisms. We describe four mechanisms for allocating decision rights, differing on two dimensions:

- **Correct Endowments (C)**: *Each agent endowed with her own decision: control of her own personal sphere* vs. **Incorrect Endowments (I)**: *Each agent endowed with other's decision: control of other's personal sphere.*
- **Trade (T):** *Agents may trade control rights.* vs. **No Trade (N):** *Agents may not trade.*

Varying along these two allows for four possible mechanisms, which we indicate by two initials above in the obvious way: CN, IN, CT, and IT.

I formally model the mechanisms as follows. After each agent privately learns her type (i.e., her preferences over red and green shirts), both agents simultaneously choose the action "trade" or "no trade". If both choose "trade", each agent makes the decision with which the other agent was initially endowed. Otherwise, each agent makes the decision with which she herself was initially endowed. In the mechanisms without trade, each agent must simply make the decision with which she was initially endowed.

Equilibrium Selection. CN and IN obviously contain unique equilibria in which each agent selects her favorite color shirt. However CT and IT may contain multiple equilibria. Each equilibrium induces a different *social arrangement* in the terminology of Section 5.3, and we could evaluate different social arrangements involving the same mechanism (i.e., game form). For simplicity, I focus on one equilibrium for each mechanism. In CT and IT, a strategy consists of a decision of whether to trade or not to trade (conditional on one's type), and then decisions about which shirt color to choose from whichever menu one has access to as a function of the first round trade decisions (again as a function of one's type). I choose to focus on equilibria that are sequentially rational. Sequential rationality implies that the agent will choose her favorite shirt color – her own or the other's – from whichever menu she faces at the second state. This leaves only the trade decisions to determine. I focus on the natural pure strategy equilibria in which each agent i selects "trade" if and only if i prefers to control the other's shirt as opposed to her own (i.e., $\gamma_i > 1$).[52] Uniqueness of equilibrium is unimportant for my purposes: My aim is to *illustrate* my proposed method of social evaluation in an interactive setting, not to come to particular conclusions about the highly stylized situation under consideration. Analysis of the equilibria is done on the assumption that each agent's behavior is dictated only by her instrumental preferences. For

[52]This means in particular that an agent's trade decision does not depend on her type (i.e., her preferred pair of shirt colors). The equilibria I focus on are are not only sequential equilibria, but also trembling hand perfect.

discussion of the possibility that agents internalize the value of intrinsic freedoms, so that these guide behavior, see Section 7.1.

Menus and Outcomes for BobInt. BobInt embodies a strong conflict: Each agent would prefer to control Ann's shirt rather than Bob's. So, if trade is allowed, it will not occur. *CN and CT will always lead to the same outcome: Each agent will always wear whichever shirt she herself prefers. IN and IT will always lead to the same outcome: Each agent will always wear whatever shirt the other agent prefers.*

An agent's equilibrium menu is the set of options from which the agent can effectively choose, given the equilibrium choice of the other. M_i^{mine} is the menu containing two lotteries each of which determine i's shirt to be a specific color (green or red), while leaving the color of $-i$'s shirt to be either red or green with equal probability. So M_i^{mine} is the menu that corresponds to controlling one's own shirt. M_i^{yours} is the menu that allows i to control the other's shirt. Define $M^{\text{choice}} = M_i^{\text{mine}} \cup M_i^{\text{yours}}$, which is the menu that allows i to choose a lottery either in menu M_i^{mine} or M_i^{yours}, whichever i prefers. That is, M^{choice} allows i to choose either agent's shirt color with the provision that whoever's shirt i does not choose is determined to be either color with equal probability.[53] Table 2 displays the equilibrium menus in each mechanism: The equilibrium menus in IN and CN are self-explanatory. In

Mechanism	**Ann's Equilibrium Menu**	**Bob's Equilibrium Menu**
CN	M_a^{mine}	M_b^{mine}
IN	M_a^{yours}	M_b^{yours}
CT	M^{choice}	M_b^{mine}
IT	M_a^{yours}	M^{choice}

Table 2: Equilibrium Menus under BobInt

CT, Bob is willing to trade, and so Ann is effectively able to choose whose shirt she will choose, so that her equilibrium menu is M^{choice}. On the other hand Ann is not willing to trade with Bob. So, his equilibrium choice is effectively M_b^{mine}. The explanation of IT is similar.

Menus and Outcomes for BothInt. In this case, the interests of the agents are more closely aligned as each agent prefers to control the other's shirt rather than her own; so an arrangement in which each controls the other's shirt satisfies both agents' preferences. In CT, trade will occur to achieve the preferred outcome. In IT trade will not occur as agents are endowed with their preferred decisions. In sum: *CN will lead each agent to wear whichever shirt she herself prefers. IN, IT, and CT will always lead to the same outcome: Each agent will always wear whatever shirt the other agent prefers.* In CT, as both agents are willing to trade, both effectively share the choice of which shirt to control, and in IT, as neither is willing to trade, neither can choose to control her own shirt.

[53] There is no subscript i on M^{choice} because it is the same menu regardless of whether i is a or b.

Mechanism	Ann's Equilibrium Menu	Bob's Equilibrium Menu
CN	M_a^{mine}	M_b^{mine}
IN	M_a^{yours}	M_b^{yours}
CT	M^{choice}	M^{choice}
IT	M_a^{yours}	M_b^{yours}

Table 3: Equilibrium Menus under BothInt

Intrinsic weights. We set the *intrinsic* weights $w_a^\circ = w_b^\circ$ in (11) so that the intrinsic freedoms of both agents are given equal weight.

Evaluating Mechanisms under BobInt. Define:

$$\begin{aligned}\Delta :&= w_b^* \left(\nu_b^* \left(M_b^{\text{yours}}\right) - \nu_b^* \left(M_b^{\text{mine}}\right)\right) - w_a^* \left(\nu_a^* \left(M_a^{\text{mine}}\right) - \nu_a^* \left(M_a^{\text{yours}}\right)\right) \\ &= w_b^* \left(\gamma_b - 1\right) - w_a^* \left(1 - \gamma_a\right)\end{aligned} \tag{14}$$

Δ is the instrumental–or consequentialist–difference between Bob's benefit if he control's Ann's personal sphere rather than his own and Ann's loss if this is so. Δ is positive if – considering only consequences – it is judged more important to grant Bob than Ann control of Ann's personal sphere (given that whoever does not control Ann's sphere, controls Bob's), and negative otherwise. Δ may be judged to be positive because Bob is judged to "care more" about controlling Ann's personal sphere.

In the following propositions, $\sim$ and $\prec$ are used to express indifference and strict preference according to the overall value function incorporating both intrinsic and instrumental values. The thresholds in the propositions depend on the value judgment about the relative importance of intrinsic as opposed to instrumental values as expressed by κ (see (11)).

Proposition 5 *Let preferences be as in BobInt. (i) Under empty priorities: CN $\sim$ CT and IN $\sim$ IT. If $\Delta < 0$, then IN $\prec$ CN, and if $\Delta > 0$, then CN $\prec$ IN. (ii) Under personal sphere priorities: CN $\sim$ CT. There exist thresholds $0 < t_0 < t_1$ such that: If $\Delta < t_0$, then IN $\prec$ IT $\prec$ CN; if $t_0 < \Delta < t_1$, then IN $\prec$ CN $\prec$ IT; and if $t_1 < \Delta$, then CN $\prec$ IN $\prec$ IT.*

The proof of Propositions 5 (and that of Proposition 6 below) involves straightforward calculations, and so is omitted.

This paragraph discusses comparisons that are independent of Δ and the next discusses those that depend on Δ. CN and CT always lead to the same outcome. Likewise, IN and IT always lead to the same outcome. When priorities are empty, procedural differences are ignored so that CN is indifferent to CT and IN is indifferent to IT. Under personal sphere priorities, CN and CT remain indifferent but IN and IT do not. This is because only IN and IT induce different freedoms *within the personal sphere*. While neither mechanism allows Ann to control her personal sphere (given prevailing behavior), IT, unlike IN, allows Bob to control his personal sphere if he chooses because Ann is willing to grant this to him (in

exchange for control over her own personal sphere). (See Table 2.) IT is preferred to IN because of the additional valuable intrinsic freedom IT grants to Bob.

In general, who should control which personal sphere depends on to whom it makes more of a difference and the magnitude of this difference (in comparison to the other normative parameters), which is captured by Δ (see (14)). Under empty priorities, Δ is all that matters, so that the *sign* of Δ determines who should control Ann's personal sphere: Bob should control Ann's personal sphere only if it matters more to him – only if $\Delta > 0$. This judgment diverges from ordinary morality. Under personal sphere priorities, the values of intrinsic freedoms have to be factored in, so that the minimum threshold t_0 required for granting control of Ann's personal sphere to Bob is greater than zero. Prevailing morality may be such that Δ never rises above the threshold. (Recall that whether Δ rises above the threshold depends on various value judgments about interpersonal comparisons of welfare as well as about the intrinsic values of freedoms.) Be that as it may, personal sphere priorities also allow us to represent a morality according to which Bob should be granted control of Ann's personal sphere, but only if he cares (or is judged to care) *enough* – where "enough" may need to be well in excess of how much Ann cares.

While the value judgments presented here do incorporate the intrinsic value of freedom, these value judgments ignore some potentially relevant aspects of freedom. Specifically they ignore the *reason* an agent has access to these freedoms. For example, control of one's personal sphere is evaluated in the same way when one is endowed with it as when another agent is willing restore it through an exchange. Future work will explore the possibility of *formally incorporating the source of constraints or reasons that freedom is restricted into the measure of freedom.*

Evaluating Mechanisms under BothInt. We now assume that preferences are as in BothInt. (14) can be equivalently be rewritten as: $\Delta = w_a^* \left(\nu_a^*(M_a^{\text{yours}}) - \nu_a^*(M_a^{\text{mine}})\right) + w_b^* \left(\nu_b^*(M_b^{\text{yours}}) - \nu_b^*(M_b^{\text{mine}})\right) = w_a^* (\gamma_a - 1) + w_b^* (\gamma_b - 1)$. This way of re-expressing Δ is more intuitive in this case, because under BothInt, both agents prefer to control the other's personal sphere, so both the differences in the sum (as rewritten) are now positive. For the same reason, Δ *must* now be positive.

Proposition 6 *Let preferences be as in BothInt. Then $\Delta > 0$. (i) Under empty priorities: $CN \prec CT \sim IN \sim IT$. (ii) Under personal sphere priorities: $IN \sim IT$. There exists a threshold t such that: if $\Delta < t$, then $IN \prec CN \prec CT$, and if $t < \Delta$, $CN \prec IN \prec CT$.*

Under BothInt, all mechanisms except CN lead to the same outcome: Each agent controls the other's personal sphere. So under empty priorities, all these mechanisms are indifferent to one another. All are consequentially superior to CN, and so under empty priorities, all are preferred to CN. For both empty and personal sphere priorities, CT is optimal because it allows agent's both maximal control, both within the personal sphere – because agents

are endowed with this control – and outside the personal sphere – because the other agent is willing to trade. So CT obviates the need to trade off freedoms against consequences as it maximally satisfies both. Under personal sphere, CT is the *uniquely* optimal mechanism (among the mechanisms we consider).

The comparison of IN and CN (or IT and CN) reveals a tension however. IN is superior in terms of outcome, but CN is superior in terms of intrinsic freedom. So there is a trade-off between consequences and freedoms; if Δ – the benefit of good consequences – is large enough, then IN is preferred, but if Δ is not sufficiently large – even if it is positive – then considerations of intrinsic freedom dominate. In the latter case, value judgments are at odds with weak Pareto: both agents prefer IN but the value system ranks CN as superior. This important issue – the conflict with Pareto – is discussed and evaluated in Section 7.4.

7 Discussion

The preceding analysis raises a number of philosophical and conceptual issues. This section discusses some of these issues.

7.1 Internalizing the Value of Freedom

An important but delicate issue concerns the degree to which agents internalize the value of intrinsic freedom.

We may think of this as either a normative or a positive question. The positive question is: *To what degree do agents internalize the the intrinsic value of freedom?* This is an empirical question. It might be posed in the form: Do agents commonly prefer choice sets that are less attractive from the standpoint of instrumental value alone? For example: An individual who practices the majority religion may be willing to sacrifice material gains or other benefits to live in a society that allows her to practice a minority religion even though she is sure she would never abandon her own religion, not because of any side-benefits, but because she wants to be free to choose her religion and not have it be chosen for her. Bartling, Fehr and Herz (2014) find experimental evidence that control is intrinsically valued in a delegation problem.[54] Psychologists have also explored the intrinsic importance that people place on autonomy and control (McClelland 1975, Ryan and Deci 1985).

However, we can also look at the question from a normative perspective by asking: *Should we value intrinsic freedoms only insofar as individuals value them?* Or should we, from a social perspective, value for example a person's freedom to vote or to practice the religion of her preference even when the person herself does not value this freedom?

The normative and positive questions are really dual. We can start by positing certain intrinsic values and ask the positive question: Do people actually tend to internalize such

[54] Bartling et al. (2014) provide further relevant references from the economic literature.

values? Alternatively, we could start with the degree to which individuals actually internalize the intrinsic value of freedom and ask the normative question: Should we only value intrinsic freedom to the degree that people value it, or should we assign a greater value to intrinsic freedom?

In the preceding analysis and its interpretation, I have assumed there to be a divergence between the intrinsic value of freedom and the internalization of this value. In part this is because this is the more interesting case to consider. In part it is because the distinctive normative force of people's claims to freedom seem to be based on the basic importance of giving people the freedom to choose rather than to be parasitic on individual preference in the same way that a value for anything might simply come from someone's preferring it. One might object that there can be no basis for valuing freedom except for people's preference to be free. Otherwise, the value of freedom must be imposed from outside, and what is the source of the moral authority to impose values? One response would be that what is good for people is not always the same as what people take to be good for themselves, but I will not pursue such a response. Another response is that if one accepts the *public good* interpretation of the value of freedom (see Section 3.4) – as at least comprising a component of this value – then one would not expect full internalization of the value of freedom; but, although I think it is important, I have not been pushing the public good interpretation in this paper. Rather, the position I take is that there is an important distinction between aggregating personal preferences and aggregating ethical views, and that the social and political foundation for accepting the intrinsic value of freedom comes from the latter source. I elaborate on this further in Section 7.5.

One can distinguish the possibility that people do not internalize the intrinsic value of their freedom *at all* from the case that people do not *fully* internalize the intrinsic value of their freedom. The interesting case that I envision is that people's ethical judgements do take intrinsic freedom – for themselves and others – to be an important value, but this ethical value is not fully reflected in behavior. However it is starker and clearer for analytical purposes, and in some respects more tractable, to focus on the former more extreme case.

In the single agent case, one might argue that the assumption that agents do not internalize intrinsic values makes no formal difference to the analysis. I could have interpreted Proposition 3 so that it is not, say, an ethical observer, but rather the agent herself who weighs and integrates her intrinsic and instrumental values. There is the further question of how we elicit the instrumental and intrinsic values separately. Such issues are largely orthogonal to the normative considerations that have been the focus of this paper. We could imagine that the agent interrogates herself about her preferences in virtue of instrumental as opposed to intrinsic considerations. In the context of a dynamic model of preference of flexibility, Ahn and Sarver (2013) propose – but do not explore in detail – the possibility that if we could observe both the choices that an agent makes *between* menus and the subse-

quent choices she makes *from* menus, we could discern whether an agent has a taste for pure freedom. Bartling et al. (2014) present an experimental design for separating (internalized) intrinsic and instrumental values.[55]

Similarly, in a interactive game theoretic setting, one might suppose that the technical analysis would be unaltered if we assumed that agents internalize the intrinsic value of freedom. Here, the issues are more subtle. In the framework of Section 5, the freedom that we evaluate is the freedom inherent in the agent's equilibrium menu, which specifies what the agent is actually able to achieve given others' behavior. Because this menu is determined by others' behavior, if an agent were to internalize the intrinsic value this would not affect her choices *from the* equilibrium menu. From her equilibrium menu, each agent would choose solely so as to optimize her instrumental interests, and the set of equilibria would remain unaltered regardless of the attitude that the agent takes to the intrinsic value of her freedom.

The above arguments – both for the single agent and the multi-agent cases – are most persuasive for interactions in which each agent only makes a single decision at a single point in time, such as when all decisions are simultaneous. However, with interconnected dynamic decisions that extend over time, there are many situations in which an agent's behavior affects her future freedom. Specifically, suppose that the agent is entitled to make a decision and she can also sell the right to make this decision. The agent has a strategy that guarantees the right to make the decision, namely, not to sell it. So from the standpoint of today, she has control over the decision in the sense that she currently has the ability to bring about whichever outcome she wishes. Still, if the agent values her intrinsic freedom, the most natural assumption is that the agent will take into account the intrinsic loss of freedom tomorrow in deciding whether to sell the decision right. My framework does not accommodate this. Indeed, the identification strategy of Bartling et al. (2014) depends on such considerations.

In sum, in both single agent and multi-agent settings, the framework of this paper best accommodates (i) situations in which an agent's current decisions do not influence her own future freedoms, both when agents do and when they do not internalize the intrinsic value of freedom. To be precise, these are situations in which at the ex ante stage, the agent (or ethical observer) contemplates a menu from which the agent will make a decision at a later interim stage that will not impact her freedoms thereafter. In contrast, the above considerations suggest that more care is required in applying the framework to (ii) situations in which an agent's current decisions may impact her own future freedoms *and* in which the agent internalizes the intrinsic value of her freedoms. Of course, the impact of a current decisions on future freedoms is a matter of degree, and so a judgment may be required as to whether the situation is better approximated by category (i) or (ii). A third type of

[55]These authors use a multi-agent interaction to identify preferences of a single decision maker.

situation is one in which (iii) an agent's current decisions impact her future freedoms but the agent does not internalize the intrinsic value of her freedoms. An instance of this is the example in Section 6. In such cases, the positive assumption that agents optimize with regard to their instrumental preferences is straightforward (or, no less straightforward than it ordinarily is). However, one might argue that the normative evaluation of the intrinsic value of freedom should be sensitive to the way that freedom evolves over time. I discuss this in the next section.

7.2 Freedom in Multistage Decisions and the Problem of Restricting One's own Freedom

When I apply the (intrinsic or hybrid) freedom measure to evaluate multistage decisions and institutions, I envision applying the freedom *only ex ante* over a lifetime or over some fixed period or interaction. The menu to be evaluated then spans the set of lotteries that the agent can induce at the ex ante stage by varying her strategy specifying future behavior following all contingencies. In particular, this means that what the agent gives up is her own responsibility and is not penalized by the freedom measure. When an agent does sacrifice some one freedom, it is typically to preserve some other freedom. As the framework studied here includes conjunctive attributes – that is, the difference between being able to exercise two freedoms simultaneously and only being able to exercise them separately (see Section 2.1)[56] – it can capture trade-offs between different freedoms from an ex ante perspective. The interesting issue arises when an agent must sacrifice a freedom with an intrinsic component for a freedom that is purely of instrumental value.

It is not entirely obvious how one should evaluate freedoms that are voluntarily dismissed by the agent herself; after all, at some point, an agent must decide, and in doing so, rule out options. Insofar as such decisions can be deemed freedom-reducing in a way that we would want our social measure to penalize (e.g., one should not be able sell oneself into slavery), this would raise interesting and complicated questions about dynamic consistency in the evaluation of freedom.[57] Moreover, an ethical observer who evaluates social freedom may not have the same attitude as an agent who internalizes her own intrinsic freedom. There might be some circumstances in which the ethical observer would have the attitude: "it is the individual's responsibility whether she gives up her freedoms; from a social standpoint,

[56] See also the closing remark of Section 2.3.

[57] It is interesting that the issue of dynamic consistency does not arise if we are concerned only with instrumental freedoms. In this case, future instrumental freedoms will automatically be folded back into the current evaluation of instrumental freedom. So, for example, if the agent is to restrict her own freedom, from an instrumental perspective, she will take the lost flexibility into account today. The asymmetry comes about from the fact that in the case of instrumental freedom the evaluation is literally derived from the agent's knowledge of what she would do in the future in various circumstances and the value of the induced outcomes, whereas in the case of intrinsic freedoms, Proposition 2 only shows that *ex ante*, it is *as if* the evaluation tracks the value of what the agent would achieve under various circumstances; at the interim stage, when those circumstances actually arrive, the agent may not conform to this as if representation.

what matters is whether the agent had the appropriate opportunity ex ante," whereas the agent could not take such a detached attitude to her future freedoms. An detailed examination of these issues must be left to future research.

7.3 Paternalism

It is natural to ask whether the hybrid measure is paternalistic. Paternalism is typically understood as a property of actions or interventions, but insofar as the hybrid measure may be action-guiding in the sense that the choice between institutions or arrangements could be guided by it, the hybrid measure can be said to inherit the property of paternalism insofar as it would recommend actions contrary to individual's preferences.

Consider the value of life example from Section 4: The hybrid measure is responsive to the agent's survival probability in a way that surpasses the agent's own concern for her survival. So *the hybrid measure disagrees with the agent on some menu rankings* (assuming no internalization of intrinsic value). An argument against the paternalism of recommendations induced by the measure is that: *The hybrid measure does not want to restrict choice, as captured by the condition* $M \subseteq M' \Rightarrow \nu(M) \leq \nu(M')$. Indeed, in the example, *the hybrid measure is more resistant than the agent to eliminating options.* Specifically, the agent is indifferent between $\left\{\left(\frac{1}{10}, w^h\right), \left(0, w^\ell\right)\right\}$ and $\left\{\left(\frac{1}{10}, w^h\right)\right\}$, but ν prefers larger menus. Also, *the hybrid measure always prefers to allow agent to choose between menus:* $\nu(M_1 \cup M_2) \geq \max\{\nu(M_1), \nu(M_2)\}$ *(and, in an interactive setting, between mechanisms).* The hybrid measure only opposes that a smaller menu is *imposed* by someone else. In this sense, it is anti-paternalistic.

The most important argument, however, may not come from the specific properties of the hybrid measure, but from more general considerations. If some authority – even with benevolent intentions – implements outcomes on the basis of the assumption that the authority knows better than the individual what is good for the individual, that is paternalistic. If a community, after deliberation and debate, decides that it will support and protect certain values for everyone, such as freedom of choice, which might not be fully internalized in the behavior of some or even all of its members, then the resulting policies may not be paternalistic.

7.4 Conflict with the Pareto Criterion

An issue closely related to paternalism is the conflict with the Pareto criterion. In the previous section, I observed that the preferences of an individual i over choice situations may not coincide with the evaluation of the hybrid measure – which also captures the value of intrinsic freedoms – for agent i. This immediately implies that it is possible that situation X Pareto dominates situation Y, but a value judgment integrating both the individuals' preferences and the intrinsic value of her freedom will prefer Y over X. This implication

follows because when there is only one agent i, X Pareto dominates Y if and only if i prefers X to Y.

Section 6, which presented the example of trade of rights to control one's personal sphere, illustrated the conflict with Pareto in a multi-agent example. When preferences satisfy the condition BothInt, then under both narrow and wide priorities, there exist parameter values such that the mechanism CN is preferred to the mechanism IN, despite the fact that IN Pareto dominates CN. This is reminiscent of the conflicts between efficiency and freedom explored by Sen (1970), Gibbard (1974), and others.

If – contrary to the treatment above – agents fully internalized the intrinsic value of their freedoms, then a modified notion of Pareto dominance that incorporated only agents' instrumental preferences would have little normative force. So the interesting case to consider is that in which the value of freedom is not derived from individual preferences for freedom, but is rather a separate self-standing value.

The conflict is interesting because the Pareto criterion may itself be thought to be founded on freedom. First, the Pareto criterion gains much of its normative force from the liberal anti-paternalistic sentiment that individuals should be sovereign over their own good. Second, there is a particularly liberal way of formulating the desirability of Pareto efficient allocations: An allocation X is Pareto efficient if and only if for each agent i, if i were allowed to choose any feasible allocation subject to the constraint that the chosen allocation would not be dispreferred to X by any other agent, i would choose X. So an allocation is Pareto efficient if and only if the only constraints (other than resource or technological constraints) on any agent's choice are conflicts with what other agents would choose. That is, there are no additional constraints stemming from the way institutions have been designed; the freedom of one agent is restricted only by the freedom of another.

This is an elegant idea, but it does not exhaust our ideals of freedom. Another aspect of the ideal of freedom is that an agent ought to have access to or priority over certain specific decisions. For example, an agent ought to have priority over her personal sphere. A Pareto efficient allocation that assigns priority to the wrong people (e.g., someone else controls my personal sphere) violates the ideal of freedom. In order for control to be legitimately transferred, the rightful owner of the decision must grant consent. *Consent* is another aspect of the ideal of freedom not captured by the Pareto criterion.

If we truly value the priorities of control inherent in freedom, then we should be willing to pay something for them. When these priorities conflict with Pareto comparisons, this payment might come in the form of sacrificing the instrumental interests honored by such comparisons. In other words, we should be willing to trade off these different aspects of freedom.

I illustrate this idea with the example of Section 6. Let us again compare CN to IN.[58]

[58]Since we are comparing CN to IN, trade has been ruled out. Fortunately, allowing trade–*from correct endowments*–would relieve the conflict, but the comparison of the suboptimal arrangements, CN and IN,

The question is whether we should *impose* the outcome preferred by the agents at the expense of their *consent* to yield control over their personal spheres. If we truly value giving agents the option of whether to yield personal control, then we should be willing to pay something to maintain the option. This means that if the net instrumental gain associated with eliminating this option is sufficiently small, we should not be willing to eliminate it. The general lesson is that even a policy that would make everyone better off might be *too intrusive* to be justified on the part of a central authority.

7.5 The Source of the Intrinsic Value of Freedom

If agents have a taste for freedom, then it is easy to understand the value judgments involving freedom: Such value judgments simply reflect the preferences of individuals and are important for the reasons that honoring any preference is important. If on the other hand, agents themselves do not have a preference for freedom in itself (but only as a means to an end), then why should we treat freedom as being important, and why should normative analysis take it seriously?

In part, this is a question in metaethics, about the foundation of moral truths. I will not engage with such metaethical questions here. However, I will say a word about the foundation of the intrinsic value of freedom in prevailing social judgments. There is an important distinction between the preferences that govern ordinary decision-making and broader value judgments. While many individuals may behave as if freedoms are only instrumentally valuable, these same individuals, or at least some of them, may share the ethical view that some basic freedoms are intrinsically important, not just for themselves, but also for others, even those who do not share this moral value. Social consensus rooted in values that are broader than personal preferences influences, among other factors, the choice of institutions in society. I view the intrinsic value of freedoms as being part of this broad social consensus.

A Appendix: Proofs of Propositions

A.1 Proof of Proposition 1

When $M \in \mathcal{M}_d$, $\{\mathcal{L}(A) : A \subseteq \bar{Z}, M \cap A \neq \emptyset\}$ is a partition of $\mathcal{A}(M)$. It follows that:

$$\nu(A) = \int_{\mathcal{A}(M)} \tilde{\lambda} d\tilde{\mu} = \sum_{A \subseteq \bar{Z}: M \cap A \neq \emptyset} \int_{\mathcal{L}(A)} \tilde{\lambda} d\tilde{\mu} = \sum_{A \subseteq \bar{Z}: M \cap A \neq \emptyset} \Lambda_d(A).$$

helps us to get clearer on the underlying value judgments. Moreover, unanimity of preference is a limiting case where interests do not even need to be traded off. The conflict between freedom and the Pareto criterion strongly suggests a conflict between freedom and welfare more generally.

For the second statement, choose a deterministic freedom measure Λ with $\Lambda(\bar{Z}) = 0$. For each $A \in \mathcal{A}_d$ with $A \neq \bar{Z}$, define $v^A = (v_z^A : z \in Z)$ by $v_z^A = \sqrt{\frac{|\bar{Z}\backslash A|}{|\bar{Z}||A|}}$ if $\delta_z \in A$ and $v_z^A = -\sqrt{\frac{|A|}{|\bar{Z}||\bar{Z}\backslash A|}}$ if $\delta_z \notin A$, where for any set B, $|B|$ is the cardinality of B. Then $v^A \in V$. Define $V^* = \{v^A : A \in \mathcal{A}^d, A \neq \bar{Z}\}$. Define λ so that $\lambda(v^A) = \Lambda(A)$, and define $\lambda(v)$ arbitrarily if $v \notin V^*$. Define $\mu(E) = |V^* \cap E|$ for all measurable $E \subseteq V$. Then $\Lambda_d^{(\lambda,\mu)}(\bar{Z}) = 0$ and for any $A \in \mathcal{A}_d$ with $A \neq \bar{Z}$,

$$\begin{aligned}\Lambda_d^{(\lambda,\mu)}(A) &= \int_{\mathcal{L}(A)} \tilde{\lambda} d\tilde{\mu} = \int_{\mathcal{L}(A)\cap\{(v,q):v\in V^*\}} \tilde{\lambda} d\tilde{\mu} = \int_{\{(v,q):v=v^A,q\in(0,1]\}} \tilde{\lambda} d\tilde{\mu} \\ &= \int_0^1 \lambda\left(v^A\right) dq \times \mu\left(\left\{v^A\right\}\right) = \Lambda(A).\end{aligned}$$

□

A.2 Proof of Proposition 2

Proof. First choose a freedom measure (λ, μ). For each $v \in V$, define $u^v = (u_z^v : z \in Z)$ by

$$u_z^v := \int_V \lambda d\mu \times \frac{v_z - \min_{z'\in Z} v_{z'}}{\max_{z'\in Z} v_{z'} - \min_{z'\in Z} v_{z'}}, \quad \forall z \in Z.$$

Observe that for all $v \in V$, $u^v \in \widehat{U}$. It is straightforward to verify that the map $v \mapsto u^v$ is one-to-one from V to $\widehat{U}$.[59] Let $p' \in \Delta(V)$ and $p \in \Delta\left(\widehat{U}\right)$ be the unique probability measures such that:

$$p\left(\{u^v : v \in E\}\right) := p'(E) := \frac{\int_E \lambda d\mu}{\int_V \lambda d\mu}, \quad \forall \text{ measurable } E \subseteq V.$$

[59] We assume here and throughout the proof that $\int_V \lambda d\mu > 0$. If $\int_V \lambda d\mu = 0$, then p can be chosen so that it puts probability 1 on $u \equiv 0$ and then $\nu(M) = \iota_p(M) = 0, \forall M \in \mathcal{M}$.

Observe that the support of p is a subset of $\{u^v : v \in V\}$ and $\lambda(v) = \int_V \lambda d\mu \times \frac{dp'}{d\mu}(v)$. Moreover:

$$\begin{aligned}
\nu(M) &= \int_{\mathcal{A}(M)} \tilde{\lambda} d\tilde{\mu} = \int_V \int_{\{q\in[0,1]:A(v,q)\cap M\neq\emptyset\}} \tilde{\lambda}(v,q)\ell(dq)\mu(dv) \\
&= \int_V \lambda(v)\ell\left(\{q \in [0,1] : A(v,q) \cap M \neq \emptyset\}\right)\mu(dv) \\
&= \int_V \lambda(v)\ell\left(\left\{q \in [0,1] : \max_{\beta\in M}(v\cdot\beta) \geq h(v,q)\right\}\right)\mu(dv) \\
&= \int_V \lambda(v)\ell\left(\left\{q \in [0,1] : \frac{\max_{\beta\in M}(v\cdot\beta) - \min_{z\in Z} v_z}{\max_{z\in Z} v_z - \min_{z\in Z} v_z} \geq q\right\}\right)\mu(dv) \qquad (15) \\
&= \int_V \lambda(v)\left[\frac{\max_{\beta\in M}(v\cdot\beta) - \min_{z\in Z} v_z}{\max_{z\in Z} v_z - \min_{z\in Z} v_z}\right]\mu(dv) \\
&= \int_U \max_{\beta\in M}(u\cdot\beta)\, p(du) \\
&= \iota_p(M).
\end{aligned}$$

Going in the other direction, note that to any element $(v,\alpha) \in V \times \mathbb{R}_{++}$, there corresponds $(\alpha[v_z - \min_{z'\in Z} v_{z'}] : z \in Z) \in \widehat{U}$ and all elements of $\widehat{U}$ (except $u \equiv 0$[60]) can be written uniquely in this way. So we can reparameterize the elements of $\widehat{U}$ in this way so that we write $\widehat{U} = V \times \mathbb{R}_{++}$. For any $u = (v,\alpha) \in \widehat{U}$, let $v^u = u$, and $\alpha^u = \alpha$ be the projections on each of the three components of u. Choose a $p \in \Delta\left(\widehat{U}\right)$. Let μ be the result of marginalizing p on V; that is $\mu(E) = p(\{u \in U : v^u \in E\})$ for all measurable $E \subseteq V$. Define $\lambda(v) = E[\alpha|v]\,(\max_{z\in Z} v_z - \min_{z\in Z} v_z)$, where $E[\alpha|v]$ is the conditional expectation of α given v. Then:

$$\begin{aligned}
\iota_p(M) &= \int_U \max_{\beta\in M}(u\cdot\beta)p(du) \\
&= \int_U \alpha^u\left(\max_{\beta\in M}(v^u\cdot\beta) - \min_{z\in Z} v_z^u\right)p(du) \\
&= \int_V E[\alpha|v]\left(\max_{\beta\in M}(v\cdot\beta) - \min_{z\in Z} v_z\right)\mu(dv) \\
&= \int_V \lambda(v)\left[\frac{\max_{\beta\in M}(v\cdot\beta) - \min_{z\in Z} v_z}{\max_{z\in Z} v_z - \min_{z\in Z} v_z}\right]\mu(dv) \\
&= \nu(M),
\end{aligned}$$

where the last equality is derived as in (15). □

[60] Since for all elements $(\alpha, v) \in V \times \mathbb{R}_{++}$ corresponding to 0 are such that $\alpha = 0$, it does not matter how the probability mass assigned to 0 is split up among them. So we ignore this detail throughout the proof.

A.3 Proof or Corollary 1

First I argue that if $\nu+c$ is a value function generated a stochastic diversity measure, then ν satisfies the properties mentioned in the corollary. Monotonicity is immediate. Proposition 2 implies that there exists $p \in \Delta(U)$ such that $\nu = \iota_p$. It is straightforward to verify that if $\nu = \iota_p$, then ν satisfies (5). That ι_p is continuous follows from Lemma S4 of Dekel, Lipman, Rustichini and Sarver (2007*b*).

Next suppose that ν satisfies the properties in the corollary. Then it is straightforward to verify that the preference relation induced by ν satisfies the axioms sufficient for an additive EU representation in Theorem 2 of Dekel, Lipman, Rustichini and Sarver (2007*a*)[61] It follows that there exists $p \in \Delta(U)$ such that:

$$\nu(M) \leq \nu(M') \Leftrightarrow \iota_p(M) \leq \iota_p(M'), \qquad \forall M, M' \in \mathcal{M}.$$

Since ν satisfies (5), and similarly, $\iota_p(\alpha M + (1-\alpha)M') = \alpha\iota_p(M) + (1-\alpha)\iota_p(M'), \forall \alpha \in (0,1), \forall M \in \mathcal{M}$, Proposition 1 of De Meyer and Mongin (1995) implies that there exists constants $b > 0$ and c' such that $\nu(M) = b\iota_p(M) + c', \forall M \in \mathcal{M}$. Proposition 2 in the current paper implies that there there exists a constant c such that $\nu + c$ is the value function generated by a diversity measure. □

A.4 Proof of Proposition 3

Proposition 2 and Corollary 1 imply that ν^* and ν° both satisfy (5). ν also satisfies (5) by assumption. The result now follows from Proposition 1 of De Meyer and Mongin (1995).

A.5 Proof of Proposition 4

I begin by defining a mixture of social arrangements more formally than I did in the text. For any game forms $\Gamma = \left((S_i)_{i\in I}, g\right)$ and $\Gamma' = \left((S'_i)_{i\in I}, g'\right)$, and $\alpha \in (0,1)$, define the game form $\alpha\Gamma + (1-\alpha)\Gamma' = \left((S''_i)_{i\in I}, g''\right)$ by $S''_i = S_i \times S'_i$ and $g\left((s_1, s'_1), \ldots, (s_n, s'_n)\right) = \alpha g(s_1, \ldots, s_n) + (1-\alpha) g(s'_1, \ldots, s'_n)$. If σ and σ' are strategy profiles in Γ and Γ' respectively, define $\left[[\sigma \times \sigma']_i(u_i)\right](s_i, s'_i) = [\sigma_i(u_i)](s_i) \times [\sigma'(u_i)](s'_i)$ and $\sigma \times \sigma' = ([\sigma \times \sigma']_i : i \in I)$. Finally define $\alpha(\Gamma, \sigma) + (1-\alpha)(\Gamma', \sigma') = (\alpha\Gamma + (1-\alpha)\Gamma', \sigma \times \sigma')$. If σ and σ' are BNEs of Γ and Γ' respectively, then $\sigma \times \sigma'$ is a BNE of $\alpha\Gamma + (1-\alpha)\Gamma'$, implying that the mixture of social arrangements is a social arrangement.

Let ν'_i be either ν^*_i or ν°_i. Choose $M, M' \in \mathcal{M}$ and $\alpha \in (0,1)$. Proposition 2 and Corollary 1 imply that

$$\nu'_i\left(\alpha M + (1-\alpha)M'\right) = \alpha\nu'_i(M) + (1-\alpha)\nu'_i\left(M'\right) \tag{16}$$

[61] A possible exception is the nontriviality axiom but this does not affect the proof.

Next, define $\hat{\nu}_i'(\Gamma,\sigma) := \nu_i'(M_i(\Gamma,\sigma_{-i}))$. The fact that $M_i\left(\alpha\Gamma+(1-\alpha)\Gamma',[\sigma\times\sigma']_{-i}\right) = \alpha M_i(\Gamma,\sigma_{-i})+(1-\alpha)M_i\left(\Gamma,\sigma'_{-i}\right)$ and (16) now imply that $\hat{\nu}_i'(\alpha(\Gamma,\sigma)+(1-\alpha)(\Gamma',\sigma')) = \alpha\hat{\nu}_i'(\Gamma,\sigma)+(1-\alpha)\hat{\nu}_i'(\Gamma',\sigma')$. The result now follows from Proposition 1 of De Meyer and Mongin (1995). □

References

Abdou, J. and Keiding, H. (1991), *Effectivity Functions in Social Choice*, Kluwer.

Ahn, D. S. and Sarver, T. (2013), 'Preference for flexibility and random choice', *Econometrica* **81**(1), 341–361.

Arrow, K. J. (1995), 'A note on freedom and flexibility', *Choice, welfare and development: a festschrift in honour of Amartya K. Sen* pp. 7–15.

Barbera, S., Bossert, W. and Pattanaik, P. K. (2004), 'Ranking sets of objects', *Handbook of Utility Theory, Vol. 2. Extensions* pp. 893–977.

Bartling, B., Fehr, E. and Herz, H. (2014), 'The intrinsic value of decision rights', *Econometrica* **82**, 2005–2039.

Bentham, J. (1782), *Of laws in general*.

Berlin, I. (1959), *Two concepts of liberty*, Clarendon.

Bernheim, B. D. and Rangel, A. (2007), 'Toward choice-theoretic foundations for behavioral welfare economics', *The American Economic Review* pp. 464–470.

Bernheim, B. D. and Rangel, A. (2009), 'Beyond revealed preference: choice-theoretic foundations for behavioral welfare economics', *Quarterly Journal of Economics* **124**(1), 51–104.

Carter, I. (1999), *A measure of freedom*, Oxford University Press.

Carter, I., Kramer, M. H. and Steiner, H. (2007), *Freedom: a philosophical anthology*, Blackwell publishing.

Dasgupta, P. (1995), *An inquiry into well-being and destitution*, Oxford University Press.

De Meyer, B. and Mongin, P. (1995), 'A note on affine aggregation', *Economics Letters* **47**(2), 177–183.

Dekel, E., Lipman, B. L. and Rustichini, A. (2001), 'Representing preferences with a unique subjective state space', *Econometrica* **69**(4), 891–934.

Dekel, E., Lipman, B. L., Rustichini, A. and Sarver, T. (2007*a*), 'Representing preferences with a unique subjective state space: A corrigendum', *Econometrica* **75**(2), 591–600.

Dekel, E., Lipman, B. L., Rustichini, A. and Sarver, T. (2007*b*), 'Supplement to 'Representing preferences with a unique subjective state space: A corrigendum"', *Econometrica Supplementary Material* **75**. http://econometricsociety.org/ecta/supmat/6232errata.pdf.

Dowding, K. and van Hees, M. (2007), 'Counterfactual success and negative freedom', *Economics and Philosophy* **23**(02), 141–162.

Dowding, K. and van Hees, M. (2009), Freedom of choice, *in* P. Anand, P. K. Pattanaik and C. Puppe, eds, 'The Handbook of Rational and Social Choice: an overview of new foundations and applications', Oxford University Press, pp. 374–392.

Foster, J. (2010), Notes on effective freedom, Technical report, Queen Elizabeth House, University of Oxford.

Gibbard, A. (1974), 'A pareto-consistent libertarian claim', *Journal of Economic Theory* **7**(4), 388–410.

Goldman, S. M. (1974), 'Flexibility and the demand for money', *Journal of Economic Theory* **9**(2), 203–222.

Green, J. R. and Hojman, D. A. (2007), 'Choice, rationality, and welfare measurement'. Working paper.

Harsanyi, J. . (1955), 'Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility', *The Journal of Political Economy* **63**(4), 309–321.

Jones, P. and Sugden, R. (1982), 'Evaluating choice', *International Review of Law and Economics* **2**(1), 47–65.

Klemisch-Ahlert, M. (1993), 'Freedom of choice', *Social Choice and Welfare* **10**(3), 189–207.

Kochov, A. (2007), 'Subjective states without the completeness axiom', *University of Rochester* .

Koopmans, T. C. (1964), On the flexibility of future preferences, *in* M. W. Shelly and G. L. Bryan, eds, 'Human Judgment and Optimality', John Wiley and Sons.

Kramer, M. H. (2003), *The quality of freedom*, Oxford University Press.

Kreps, D. M. (1979), 'A representation theorem for "preference for flexibility"', *Econometrica* **47**, 565–577.

Lowry, R. and Peterson, M. (2011), 'Cost-benefit analysis and non-utilitarian ethics', *Politics, Philosophy & Economics* pp. 1–22.

McClelland, D. C. (1975), *Power: The inner experience.*, Irvington.

Mill, J. S. (1859), *On liberty.*

Moulin, H. and Peleg, B. (1982), 'Cores of effectivity functions and implementation theory', *Journal of Mathematical Economics* **10**(1), 115–145.

Nehring, K. (1999), 'Preference for flexibility in a Savage framework', *Econometrica* **67**(1), 101–119.

Nehring, K. and Puppe, C. (2002), 'A theory of diversity', *Econometrica* **70**(3), 1155–1198.

Nehring, K. and Puppe, C. (2008), 'Diversity and the metric of opportunity'. Working paper.

Nussbaum, M. C. (2011), *Creating capabilities*, Harvard University Press.

Pattanaik, P. K. and Xu, Y. (1990), On ranking opportunity sets in terms of freedom of choice, Technical report, Université catholique de Louvain, Institut de Recherches Economiques et Sociales (IRES).

Peleg, B. (1984), *Game theoretic analysis of voting in committees*, Cambridge University Press.

Peleg, B. and Peters, H. (2010), *Strategic social choice: stable representations of constitutions*, Springer.

Raz, J. (1991), 'Free expression and personal identification', *Oxford Journal of Legal Studies* **11**, 303–324.

Rommeswinkel, H. (2014), 'Measuring freedom in games'. Working paper.

Rubinstein, A. and Salant, Y. (2012), 'Eliciting welfare preferences from behavioural data sets', *The Review of Economic Studies* **79**(1), 375–387.

Ryan, R. M. and Deci, E. (1985), *Intrinsic motivation and self-determination in human behavior*, Plenum.

Sen, A. (1970), 'The impossibility of a paretian liberal', *The journal of political economy* pp. 152–157.

Sen, A. (1979), 'Equality of what?'. Stanford University: Tanner Lectures on Human Values.

Sen, A. (1999), *Development as freedom*, Oxford University Press.

Sen, A. (2009), *The idea of justice*, Harvard University Press.

Sher, I. (2014), 'Freedom as control'. Working paper.

Taylor, C. (1979), What's wrong with negative liberty, *in* A. Ryan, ed., 'The Idea of Freedom', Oxford University Press, pp. 175–194.

Thaler, R. H. and Sunstein, C. R. (2003), 'Libertarian paternalism', *American Economic Review* pp. 175–179.

Thaler, R. H. and Sunstein, C. R. (2008), *Nudge: Improving decisions about health, wealth, and happiness*, Yale University Press.

Van Hees, M. (1999), 'Liberalism, efficiency, and stability: some possibility results', *Journal of Economic Theory* **88**(2), 294–309.

Van Hees, M. (2000), *Legal reductionism and freedom*, Kluwer Academic Publishers.

Weitzman, M. L. (1992), 'On diversity', *The Quarterly Journal of Economics* pp. 363–405.

Weitzman, M. L. (1998), 'The Noah's ark problem', *Econometrica* pp. 1279–1298.

Zamir, E. and Medina, B. (2008), 'Law, morality, and economics: Integrating moral constraints with economic analysis of law', *California Law Review* pp. 323–391.

Zamir, E. and Medina, B. (2010), *Law, economics, and morality*, Oxford University Press.