Journal of Economic Behavior & Organization 125 (2016) 148–161

Contents lists available at ScienceDirect

Journal of Economic Behavior & Organization

journal homepage: www.elsevier.com/locate/jebo

A representation theorem for guilt aversion☆

Martin Kaae Jensen[a,*], Maria Kozlovskaya[a,b]

[a] Department of Economics, University of Leicester, Leicester LE1 7RH, UK
[b] The Business School, University of Huddersfield, West Yorkshire HD1 3DH, UK

ARTICLE INFO

Article history:
Received 31 July 2015
Received in revised form 31 January 2016
Accepted 11 February 2016
Available online 5 March 2016

JEL classification:
C72
C91
D03

Keywords:
Guilt aversion
Utility representation
Psychological games

ABSTRACT

Guilt aversion has been shown to play an important role in economic decision-making. In this paper, we take an axiomatic approach to guilt by deducing a utility representation from a list of easily interpretable assumptions on an agent's preferences. It turns out that our logarithmic representation can mitigate the problem of multiplicity of equilibria to which psychological games are prone. We apply the model in three well-known games and show that its predictions are consistent with experimental observations.

1. Introduction

Guilt is the experience of discomfort that follows when we violate a personal or social standard. If an action raises income but disappoints our own or other individuals' expectations of us, it may trigger our guilty conscience. Any individual who is sufficiently averse to this discomfort may therefore refrain from taking the action in the first place. Guilt aversion is able to explain a vast array of behaviors, including cooperation (Miettinen and Suetens, 2008), altruism (Andreoni and Rao, 2011), conformism (Khalmetski, 2015), group favoritism (Güth et al., 2009) and reciprocity (Chang et al., 2011), and economic experiments indicate that it is indeed an important determinant in a variety of different situations (Ketelaar and Au, 2003; Charness and Dufwenberg, 2006; Hopfensitz and Reuben, 2009; Geng et al., 2011; Battigalli et al., 2013). More recently, guilt averse behavior has also made its way into macroeconomic modeling. Thus Ahrens and Snower (2014) incorporates guilt and envy into a dynamic stochastic equilibrium model and shows that when these emotions are experienced by workers, a Phillips curve relationship between inflation and output can be generated.

A popular way to model emotions, including guilt, is to include them as inputs in agents' utility functions. Particularly important for experimental work are linear utility representations with money and guilt as the inputs (Battigalli and

☆ We would like to thank Kirill Borusyak, Subir Bose, Martin Dufwenberg, David Rojo Arjona, Rajiv Sarin, conference participants at the University of Ghent and the University of Manchester, as well as the editor William S. Neilson and two anonymous referees for their helpful comments and constructive suggestions. Maria Kozlovskaya gratefully acknowledges financial support from the ESRC.

* Corresponding author. Tel.: +44 116 252 2901.
E-mail addresses: mj182@leicester.ac.uk (M.K. Jensen), m.kozlovskaya@hud.ac.uk (M. Kozlovskaya).

http://dx.doi.org/10.1016/j.jebo.2016.02.001
0167-2681/

Dufwenberg, 2007; López-Pérez, 2010; Chang et al., 2011; Battigalli et al., 2013; Miettinen, 2013; Khalmetski et al., 2015).[1] This paper's main theoretical objective is to axiomatize utility representations of guilt-averse preferences. Specifically, axioms are presented that are necessary and sufficient for (i) the linear representation mentioned a moment ago, (ii) a representation that is logarithmic in money and linear in guilt, and (iii) a general additively separable utility representation of money and guilt. Call the sacrifice ratio between money and guilt ("how much money an agent is willing to pay to lower guilt by one unit") the *price of a clear conscience* (PCC). For well-behaved preferences we find that (i) obtains if and only if the PCC is constant for all money-guilt combinations; (ii) holds if and only if for any two levels of income the relative PCC equals the relative income, and (iii) derives whenever a suitably redefined "double cancellation condition" (Debreu, 1960) is satisfied.

By tracing specific utility representations to the level of preferences, we are able to shed light on the deeper psychological conditions that they entail vis-a-vis the previously mentioned personal or social standards. In doing so, we quickly end up concluding that the assumptions about an agent's moral compass embodied in (i) are problematic. While (iii) is not subject to this critique, it has — as will become clear from the following discussion — too many degrees of freedom to provide a useful alternative in strategic settings. This motivates our introduction of (ii) as the simplest realistic alternative — and it is important to stress, this is *not* an ad hoc alternative but one grounded in moral/psychological considerations. With this in hand we then — in what is arguably the paper's main contribution to existing literature — reanalyze a number of famous laboratory games, namely the Dictator game, the Public Good Provision game, and the Prisoners' Dilemma. This exercise provides further support for model (ii), but we postpone the specifics until Section 4.

To the best of our knowledge, the only existing paper concerned with the axiomatization of guilt-income representations is López-Pérez (2010). López-Pérez (2010) proposes a utility function exhibiting guilt aversion and provides axiomatic foundation for it. The study also features a discussion of the psychological foundation of guilt and shame and links the feeling of guilt to internalization of a social norm. The paper differs from ours in a number of ways, however, most importantly in the definition of guilt. In López-Pérez (2010), guilt is binary (-1 if the social norm was breached and 0 otherwise), and for the value of guilt to be determined, an exogenous social norm must be specified. By contrast, in our setting guilt is a real number with the standard interpretation as the difference between an opponent's actual and expected payoff (see *e.g.* Battigalli and Dufwenberg (2007) as well as the discussion in Section 2). Finally, the properties of the preference relation in López-Pérez (2010) depend on what other players do, hence any given representation is only defined within a specific game. By contrast, our preference relation is set on an abstract guilt-money space and thus can be applied to both decision and game theory. An axiomatic approach to the broader field of other-regarding preferences has been pursued by several authors, most notably Neilson (2006) and Sandbu (2008). Both papers axiomatize general function forms: Additively separable reference-dependent utility in the former and CES-utility in the latter. What sets these studies apart from the results of the current paper is our focus on specific functional forms with few enough free parameters to be testable in the laboratory.

The rest of the paper proceeds as follows. Section 2 introduces money-guilt utility functions. Section 3 develops a theory of moral choice and presents our axiomatization results. Section 4 studies the experimental evidence in the three games mentioned above as well as further discussion. The Appendix contains proofs.

2. Existing literature and the logarithmic alternative

The first formal model of guilt aversion was proposed by Battigalli and Dufwenberg (2007). They define guilt as the perceived payoff loss inflicted on another player, *i.e.*, as the difference between an opponent's expected payoff $E(m_j)$ and actual payoff m_j:

$$G(m_j, E(m_j)) = \max\{0, E(m_j) - m_j\}. \tag{1}$$

To be precise, since a player i does not know exactly how much his opponent j expects, $E(m_j)$ is i's belief about j's expectation. That makes guilt, and a guilt-averse agent's utility, a function of second-order beliefs (cf. Geanakoplos et al. (1989), Attanasi and Nagel (2007), Battigalli and Dufwenberg (2009)). Battigalli and Dufwenberg (2007) also propose a utility function over money and guilt (2), which has been extensively used in subsequent theoretical and experimental research.[2]

$$u_i(m_i, G) = m_i - \theta G. \tag{2}$$

Here m_i is the decision-maker's monetary payoff, G is the guilt he experiences, and θ is a guilt sensitivity parameter. A key advantage of such an approach is that it endogenizes the reference point $E(m_j)$ which with a formulation such as (2) is implicitly solved for in equilibrium. A constant marginal rate of substitution (MRS) between money and guilt arguably has a drawback, however: It can explain nearly any observed behavior. To illustrate with an often studied example, consider the so-called Dictator game, in which one player (the Dictator, hereafter D) decides upon the division of the total endowment T between himself and the other player (the Recipient, hereafter R). His donation to R, m_R, is hence his strategy. In Psychological

[1] Miettinen (2013) considers a linear utility over money and guilt in the main text of his paper. In the Appendix he studies a more general function with a weakly convex guilt component, which he adopts for technical convenience but finds difficult to justify. The alternative functional form proposed in this paper also implies convex preferences over money and guilt, but is grounded in deep psychological considerations.

[2] Examples are Battigalli and Dufwenberg (2009), Chang et al. (2011), Battigalli et al. (2013), Khalmetski (2015).