ERASMUS JOURNAL FOR PHILOSOPHY AND ECONOMICS
VOLUME 11, ISSUE 1, SPRING 2018

The Erasmus Journal for Philosophy and Economics (EJPE) is a peer-reviewed bi-annual academic journal supported by the Erasmus Institute for Philosophy and Economics, Faculty of Philosophy, Erasmus University Rotterdam. EJPE publishes research on methodology of economics, history of economic thought, ethics and economics, and the conceptual analysis of inter-disciplinary work relating economics to other fields. EJPE is an open-access journal. For additional information, see our website: <http://ejpe.org>.

EDITORS

Huub Brouwer
Osman Çağlar Dede
Willem van der Deijl
Erwin Dekker
Vaios Koliofotis

ADVISORY BOARD

Erik Angner, Roger Backhouse, Constanze Binder, Marcel Boumans, Richard Bradley, Nancy D. Cartwright, Christopher Clarke, David Colander, John B. Davis, Sheila C. Dow, Francesco Guala, Till Grüne-Yanoff, D. Wade Hands, Conrad Heilmann, Frank Hindriks, Geoffrey Hodgson, Elias Khalil, Arjo Klamer, Alessandro Lanteri, Aki Lehtinen, Uskali Mäki, Caterina Marchionni, Deirdre N. McCloskey, Serena Olsaretti, Mozaffar Qizilbash, Julian Reiss, Ingrid Robeyns, Malcolm Rutherford, Margaret Schabas, Eric Schliesser, Esther-Mirjam Sent, Robert Sugden, Alex Voorhoeve, Jack Vromen.

ISSN: 1876-9098

ERASMUS JOURNAL FOR PHILOSOPHY AND ECONOMICS

VOLUME 11, ISSUE 1, SPRING 2018

TABLE OF CONTENTS

ARTICLES

SPECIAL CONTRIBUTION

BOOK REVIEWS

PHD THESIS SUMMARIES

Erasmus Journal for Philosophy and Economics,
Volume 11, Issue 1,
Spring 2018, pp. 1-30.
https://doi.org/10.23941/ejpe.v11i1.284

Hybrid Vigor: Coherence and Correspondence Criteria for Heuristics

PATRICIA RICH
University of Hamburg

Abstract: The ecological approach to rationality involves evaluating choice processes instead of choices themselves, and there are good reasons for doing this. Proponents of the ecological approach insist that objective performance criteria (such as monetary gains) replace axiomatic criteria, but this claim is highly contentious. This paper investigates these issues through a case study: 12 risky choice processes are simulated, and their performance records are compared. The first criterion is conformity to the Expected Utility axioms; the Priority Heuristic stands out for frequently violating Transitivity. Next, the Expected Value criterion is applied. Minimax performs especially poorly—despite never violating an axiom—highlighting the tension between axiomatic (coherence) and objective (correspondence) criteria. Finally, I show that axiom violations carry high costs in terms of expected value. Accordingly, coherence does not guarantee objectively high performance, but incoherence does guarantee diminished performance.

Keywords: ecological rationality, expected utility, transitivity, independence, priority heuristic, simulations.

JEL Classification: A12, B40, D81

1. INTRODUCTION

When it comes to the task of judging whether an agent's choices are rational, two approaches vie for dominance. The first and traditional method is to apply the Expected Utility (EU) axioms to see whether a choice pattern is *coherent* (Mas-Colell et al. 1995, Chapter 6). This is a

AUTHOR'S NOTE: The author gratefully acknowledges her Ph.D. committee at Carnegie Mellon, audiences at the 2016 SAET conference and 2017 EIPE Anniversary conference, colleagues in the Entscheidung, Wissen, Wissenschaft seminar at the University of Hamburg, three anonymous referees and the editor for valuable feedback at various stages of this project. She also acknowledges funding from the ERC project 'Epistemic Utility Theory' and DFG project 'Knowledge and Decision' while this work was carried out.

compelling method because (1) it accommodates the subjective aspect of choice goodness (via *utility*), but (2) the axiomatic test is applied to observed choices and is therefore empirically grounded, and (3) the axioms themselves are intuitive, sensible, and bolstered by many proofs and arguments (see, for instance, Gilboa 2009, Chapter 6). The second approach is advocated by many psychologists under the banner of Ecological Rationality (ER). Proponents of ER forcefully criticize EU, proposing that processes should be assessed relative to particular contexts of application (see, for example, Gigerenzer and Selten 1999, Gigerenzer et al. 2011). ER focuses especially on simple decision and inference heuristics, but most important is that the processes can be precisely described by a series of easily-programmable steps. Processes are then judged on objective scales according to how fast, frugal, and accurate they are. This approach is also quite compelling, but for different reasons: firstly because it addresses how and why people make the choices they do, secondly because it is conducive to the project of improving people's choices, and thirdly because objective success is undeniably important.

EU rationality is a *coherence* standard because it checks that choices fit together in a particular way, as captured by the EU axioms. In contrast, the objective standards that ER advocates are *correspondence* standards.[1] The coherence/correspondence divide is now central to the debate about rationality standards. The debate between the approaches is persistent, with no agreement about whether or how they might be reconciled (Wallin 2013, Berg and Gigerenzer 2006, Sturm 2012; see also Rich 2016 for more extensive discussion and detailed literature references).

This paper presents a case study to support a methodological claim about how rationality should be evaluated, in light of this debate. I propose that we should evaluate processes using a hybrid method, simulating them and applying both the relevant axioms (here, the EU axioms) and the relevant objective standards (here, wealth) to the results. Doing so, I claim, retains the advantages of both EU and ER; the method has *hybrid vigor*, just as hybrid organisms are often more robust than either parent due to their increased genetic diversity. In the present case study, I simulate choice heuristics to choose between

[1] The distinction has a long tradition in philosophy but was brought into the rational choice discussion by Hastie and Rasinski (1988); see also Berg et al. (2016, 190) and Hammond (1996, 2007).

lotteries and show that the hybrid method yields more satisfying rationality assessments than either EU or ER on its own. I have defended the basic methodology elsewhere—on both theoretical (Rich 2016), and formal (Rich 2018) grounds—and the present case study serves as a proof of concept, as both EU and ER are informative regarding the heuristics in question.

The paper is structured as follows. Section 2 provides the conceptual background of the case study, justifying each step of my analysis and explaining how the steps combine to support the hybrid method. Section 3 describes the heuristics and the lotteries. Section 4 compares the heuristics' conformity with the EU axioms. Section 5 compares the heuristics using the objective criterion of Expected Value (EV). These sections, therefore, apply a coherence and a correspondence criterion, respectively. Taken together, the most frequent axiom violators tend to leave more money on the table, but the reverse is not true. This highlights the tension between coherence and correspondence criteria. Section 6 goes on to show that the more frequent axiom violators tend to leave money on the table because incoherence and objective losses coincide in a strong sense: axiom violations are associated with significant foregone profit, over 30% in this context. Section 7 discusses both the methodological and the practical implications of the case study.

2. MOTIVATING THE THREE-PART HYBRID METHOD

2.1. Expected Utility for Processes

A superficial difference between EU and ER is that EU evaluates choice patterns, whereas ER evaluates processes. Although processes and EU are seldom combined, there is no principled reason why they should not be. It is often more useful to evaluate processes; for example, teaching people *how* to choose is more efficient than teaching them *what* to choose in each case. I, therefore, adopt a process-based approach—as recommended by ER—and hereafter leave this point implicit.

This section motivates each part of the case study in turn. For present purposes, the best way to motivate the use of EU is to explain a bit of its history. It is designed to circumvent a particular problem—namely that preference is subjective and not directly observable—that plagues more direct approaches.

The development of modern EU theory involved two key steps, first from objective value to subjective utility, and second from free-floating utility to utility grounded in preferences. In the early days of mathematical decision theory, the value of a gamble was taken to be its EV, that is, the sum of the possible outcomes, each weighted by its probability. There is a serious problem with this theory, namely that it assumes that every additional dollar is equally valuable to the agent. This assumption is false, though: people typically have diminishing marginal utility for money (also known as risk aversion). The inadequacy of EV was revealed by the well-known St. Petersburg Paradox, and Daniel Bernoulli (1853 [1738]) dissolved this paradox by explaining how value and utility could come apart, thus taking the first key step in the development of EU theory.

Although the notion of utility is intuitive, Bernoulli's version was not sufficiently scientific because it was simply posited as a quantity. Von Neumann and Morgenstern (1944, Chapter 1) solved this problem by providing a set of axioms such that an agent whose preferences satisfy those axioms is provably representable as maximizing a numerical utility function, while an agent who violates any of the axioms cannot be so represented. Their work allows the utility function to be inferred from choice data. This was the second critical step in the development of EU theory. Fishburn (1989) is an excellent historical reference on this topic with many pointers to further literature.

The point of this history is not to claim that EU theory is the best or the only way to evaluate choices. Rather, the point is that EU theory incorporates an absolutely crucial insight (that utility may legitimately differ from objective value) and solves a difficult problem (that of inferring how an agent actually values the options at hand). The insight cannot be ignored; at most it could be argued that value approximates utility well enough in some restricted context. Similarly, rejecting the axiomatic solution would require addressing the problem of inferring utility in some other way.

With this justification in place, the case study starts by comparing the heuristics using the EU axioms. The idea is that the more often a heuristic produces an axiom violation, the worse its choices are. This is because axiom-violating choices are guaranteed to be suboptimal from the agent's perspective.

The modern formulation of von Neumann and Morgenstern's theory, as found in Mas-Colell et al. (1995, Chapter 6), includes two axioms with implications for which choice patterns are rational: Transitivity and

Independence. Let $X \succ Y$ stand for "Lottery X is chosen over lottery Y." Let A, B, and C be arbitrary lotteries. Then we have the following axiom:

Transitivity If $A \succ B$ and $B \succ C$ then $A \succ C$

Lotteries can also be compounded by applying a probability distribution to a set of lotteries to yield a new lottery; for example, given any probability p, we can define a compound lottery $(p \cdot A\ ;\ (1-p) \cdot C)$ which gives lottery A with probability p and lottery C otherwise. Then we have the axiom:

Independence If $A \succ B$ then $(p \cdot A\ ;\ (1-p) \cdot \mathrm{C}) \succ (p \cdot B\ ;\ (1-p) \cdot \mathrm{C})$

For example, suppose that I choose \$5 over a coin flip between \$0 and \$10. Then I am offered a choice between a coin flip that pays nothing or \$5, and a coin flip that pays nothing or even chances of \$0 or \$10. Independence implies that I choose the first option because the new initial .5 chance of getting nothing is common to both options and should not reverse my initial preference.

2.2 Questioning Coherence: Ecological Rationality

Both Transitivity and Independence are taken to be normative because an agent who violates them chooses *incoherently*, as the choices seem to contradict each other. The real-world relevance of coherence has been questioned, however, and some recent criticisms come from proponents of ER. For example, Berg (2014) argues that,

> [w]ithout the link from conformity with an axiomatized rationality to an external performance metric, these rankings in the hierarchy of rationalities may not be normatively relevant (380).

Then, he suggests,

> If the compelling normative principle is, for example, wealth, then why not simply study the correlates of high-wealth-producing decision procedures and rank those procedures according to the wealth they produce? (382).

This criticism is especially valuable because it combines skepticism about coherence with an alternative proposal, namely that we use correspondence standards as relatively direct measures of performance.

Berg cites “health, wealth, and happiness” as relevant measures of performance and suggests that we determine which heuristics are best by figuring out, essentially, which heuristics tend to yield more of specific goods that people value (with wealth being the natural candidate for lottery choices).

The foregoing makes clear two difficulties with Athis suggestion: the wealth standard may differ dramatically from the subjective standard of utility, and we lack a good way of measuring utility without using the EU axioms. The correspondence standard is therefore imperfect, but the coherence standard is imperfect too. One weakness is that, while an axiom violation indicates a suboptimal choice, more information would be needed to say how much worse the chosen option is. A more pressing concern is that EU’s coherence test is very weak; a violation proves the agent *cannot* be (represented as) a utility maximizer, but it is never possible to prove that the agent is, *in fact*, maximizing utility. A perfectly coherent agent might achieve terrible health outcomes and make very little money. A single choice is always coherent. Given general facts about human psychology, then, the coherence standard does not capture everything that matters, and a correspondence standard would be a valuable supplement.

The case study, therefore, implements Berg’s suggestion to evaluate heuristics based on the wealth they would produce. The most reasonable way of carrying out this proposal is to compare the EVs of the heuristics’ choices, and this is the primary standard that I use. The EV metric is discussed further in Section 5.1.

2.3 Connecting Coherence and Correspondence

Given that the first two parts of my case study (motivated in sections 2.1 and 2.2) rank the heuristics according to EU and EV, the natural third step is to determine how these two standards are related, and thereby to answer the question of whether coherence is linked with objective success in a concrete application. ER proponents have suggested a negative answer in general. Berg (2014) points out a lack of evidence that real agents who violate EU will fare badly in an objective sense. In the same spirit, Arkes and colleagues survey the empirical literature, emphasize a lack of evidence that incoherent choices are costly, and decry “the widespread assumption that coherence is a universal, domain general criterion of rationality” (2016, 31). These points mirror the familiar criticism of Dutch Book arguments (see Hájek 2009 for a survey): why should we think that a real-life incoherent agent would have their incoherence exploited, or that they would lose a lot of money before they realized what was happening? The source of skepticism about coherence, then, is a lack of proof that it is strongly correlated with real-world success. The lack of proof may simply be the consequence of proof not being sought, however, and this paper's results suggest that this is the case.

While it is considered an open question whether incoherent choices are indeed bad for the chooser, it is widely recognized that coherent choices need not be good in any objective sense; for example, a Brain in a Vat may be fully coherent but is arguably wrong about everything. So, to gather the right evidence regarding the connection between coherence and correspondence, I directly measure the cost of incoherent choices in the third part of the case study. The striking result is that incoherent choices are not only objectively worse than coherent choices, but dramatically so, yielding about a third less value on average.

2.4 Related Simulation Studies

We are now in a good position to situate the present paper with respect to related work. Especially important are papers by Thorngate (1980), Johnson and Payne (1985), and Bordley (1985), which compare the performance of heuristics by simulating their lottery choices. This work can be seen as a precursor to ER in several ways: it focuses on simple heuristics, prioritizes efficient decision-making over optimal performance, addresses the importance of context in determining how well a heuristic performs and employs correspondence performance criteria. Several of the heuristics studied in this paper appear in those earlier papers.

While these earlier authors recognize that choices would ideally be compared according to their subjective utilities, they also recognize that utilities differ across agents and contexts and must be inferred from behavior. Hence, as Johnson and Payne (1985, 396-397) explain, they use EV as a substitute (just as I do in Section 5.1; see also Bordley 1985, 234). Responsibility is left to the agents to choose heuristics that suit them; the authors aim only to enable informed choices (I return to this point in Section 7.2).

Nonetheless, the drawbacks of the EV approach are clear. Descriptively, we know that agents are usually not best described as having linear utility for money; instead, the so-called "fourfold pattern" of risk preferences enjoys strong support (Markowitz 1952, Tversky and Kahneman 1992). Normatively, risk-aversion and risk-affinity are widely considered rationally permissible but are ruled out by EV. The problem is that a heuristic may perform poorly according to the EV criterion, even though it would allow agents to satisfy their preferences efficiently; it would be hard to tell whether this was the case from EV data alone. With respect to previous simulations, then, this paper is novel in that it uses an axiomatic performance standard to circumvent this problem. I present the simulation set-up in the next section.

3. Simulation Set-Up

3.1 The heuristics

I compare twelve heuristics, of which seven come from traditional decision theory (and of these, five are variants of the Hurwicz Criterion), four come from earlier simulation papers, and the last was developed by ER. Each heuristic takes two lotteries as input. Here are the heuristics and their definitions:

Minimax Choose the lottery with the greater minimum payoff. Be indifferent if the minima are equal.

Maximax Choose the lottery with the greater maximum payoff. Be indifferent if the maxima are equal.

Hurwicz$_\alpha$ For each lottery, multiply the minimum gain by α and the maximum gain by $(1- \alpha)$. Sum these products to get the lottery's Hurwicz$_\alpha$ value. Choose the lottery with the greater Hurwicz$_\alpha$ value, or be indifferent if they are equal. In general, $\alpha \in [0,1]$. Here, $\alpha \in \{.1,.25,.5,.75,.9\}$.

Equiprobable Average the outcome values for each lottery. Choose the lottery with the greater average. Be indifferent if these are equal.[2]

Probable Define the 'probable' outcomes for a given lottery as those outcomes with a probability of at least (1/the number of outcomes). Choose the lottery with the greater average of these 'probable' outcomes, and be indifferent when these averages are equal.

Least Likely Choose the lottery with the smaller probability attached to its minimum payoff. Be indifferent when these are equal.

Most Likely Choose the lottery whose most probable outcome is greater; use the average when there are multiple outcomes with

[2] This is equivalent to treating all outcomes as equiprobable, as the name suggests.

maximal probability. When this quantity is equal for two lotteries, be indifferent between them.

Priority Heuristic If the difference between the minimum gains of the lotteries differs by at least 10% of their maximum gain, choose the lottery with the greater minimum. Else, if the probabilities attached to these minima differ by at least .1, choose the lottery with the smaller probability of getting the minimum. Else, if the lotteries' maximum gains differ by at least 10% of the overall maximum gain, choose the lottery with the greater maximum. Else, choose the lottery with the higher probability attached to its maximum gain. Be indifferent if these are equal.

The first three heuristics are old staples in decision theory, proposed for decision under uncertainty (when the outcome probabilities are unknown). As such, they ignore probabilities entirely. Minimax simply chooses the lottery with the best worst-case outcome, while Maximax chooses the lottery with the best best-case outcome. These are limiting cases of the Hurwicz Criterion, which compromises by assigning weight $\alpha \in [0,1]$ to a lottery's worst outcome and $\beta = 1 - \alpha$ to its best outcome, choosing the lottery with the greater weighted sum. Since Hurwicz is characterized by its parameters in this way, I test five different Hurwicz criteria, with $\alpha \in \{.1, .25, .5, .75, .9\}$. Intuitively, then, these heuristics cover the spectrum from extreme caution or pessimism to extreme risk-affinity or optimism. (See Luce and Raiffa 1957, Chapter 13.2 for an overview.)

Equiprobable, Probable, Most Likely, and Least Likely all appear in earlier simulations (Thorngate 1980, Johnson and Payne 1985, Bordley 1985).[3] Note, however, that Equiprobable is essentially the "principle of insufficient reason" that appears in the literature on uncertainty (Luce and Raiffa 1957, Chapter 13.2). Like the other heuristics, these ignore a lot of available information: Equiprobable ignores probabilities entirely, while the others use probabilities (outcomes) in a limited way to determine which outcomes (probabilities) to attend to.

The Priority Heuristic (PH) is of special interest as one of the hallmarks of the ER program. While its precise thresholds are an

[3] This paper studies pairwise choices between lotteries with all non-negative outcomes, since the PH is designed for such choices and they are the most commonly studied. Applicability to this task rules out some heuristics studied elsewhere, such as Tversky's 'Elimination by Aspects' (Tversky 1972) and the 'Better-than-Average heuristic' (Thorngate 1980).

idealization facilitating implementation, it is psychologically quite plausible. Its creators argue that it provides a compelling explanation for many observed patterns in lottery choice, including the paradoxical Allais pattern, the 'fourfold pattern of risk', and the 'Certainty' and 'Possibility' effects (Brandstätter et al. 2006).

To illustrate the heuristic, let us see how it reproduces the typical response pattern in the Allais situation. The first choice is between the lotteries we will call *A* and *B*:

A	$1 million for sure
B	$1 million with probability .89, $5 million with probability .10, and nothing with probability .01

The second choice is between the lotteries referred to as *C* and *D*:

C	$1 million with probability .11 and nothing with probability .89
D	$5 million with probability .10 and nothing with probability .90

The PH is lexicographic, which means that it considers a series of possible reasons for choice, in order, until one of those reasons is decisive. The first of the PH's reasons is the minimum gain. This reason decides in favor of Lottery *A* over Lottery *B* in the Allais case: *A*'s worst outcome of $1 million is compared to *B*'s worst outcome of nothing; this difference exceeds 10% of $5 million, so *A* is chosen. (Put simply, *A* guarantees a good outcome.) In contrast, when the PH compares lotteries *C* and *D*, the lotteries have the same worst outcomes (nothing), and so this reason is not decisive. The probabilities of the minima are compared next, but these are too similar (.89 vs. .90). The PH, therefore, compares the lotteries' maxima, and Lottery *D* is chosen because its maximum ($5 million) is sufficiently large in comparison to *C*'s ($1 million).

Although the PH is intended as a descriptive heuristic, ER proponents repeat that the program's goals are threefold, "descriptive, normative, and engineering" (see, for example, Gigerenzer et al. 2011, xix), so it is fair to subject the PH to normative appraisal. Nonetheless, I am not committed to the PH as the correct explanation. Regardless of its descriptive status, the PH is of interest because it is prominent in the literature and provides a useful comparison between lexicographic and one-step heuristics. Having described the heuristics, I turn next to the lotteries they choose between.

3.2 The Lotteries

A formal lottery precisely represents a risky option; the possible outcomes (here given in dollar values) are listed along with the objective probability with which each outcome occurs (recall the Allais lotteries in Section 3.1). The heuristics are tested on lotteries appearing in the decision science literature, specifically Allais (1953), Brandstätter et al. (2006), Binmore (2009, 50-52), and Kahneman and Tversky (1979). Additionally, around 45% of the lotteries are randomly generated and come from the Technion Prediction Tournament (2008), a competition between algorithms to best predict human lottery choices. These sources provide an initial set of 171 unique lotteries. The heuristics are simulated to make choices on every pair of lotteries from this initial set.[4]

The main requirement for this test set is that it be sufficiently large and diverse to ensure that the results are not an artifact of some feature of the test set that a broader sample of lotteries would not share. The lotteries are diverse in terms of the possible outcomes: Most (including most Tournament lotteries) have two possible outcomes, but some have more, and quite a few have five; some also offer a particular outcome for certain. The outcomes vary from \$0 to millions, with the entire range from \$0 to thousands well represented (the Tournament lotteries generally have lower values). The fact that all outcomes are non-negative simplifies the analysis at little cost; for example, the loss version of the PH perfectly mirrors the gain version so that replacing all gains (\$*x*) with their negation (−\$*x*) would not influence the results.[5]

The lotteries also cover a broad spectrum of within-lottery outcome variances: low-variance lotteries are especially prevalent (a natural consequence of including riskless options), but relatively high variance lotteries are well-represented and the range in between is covered. Similarly, a disproportionate number of lotteries have no variance in the probability distributions (as when there is a sure payoff or outcomes are equiprobable), but apart from this, the representation of the range of possible variances is roughly even.

[4] Upon request, the author can share the lottery tables, the spreadsheets used to produce the choices, the code used to analyze them, and so forth.

[5] While mixed lotteries—those in which both losses and gains are possible—are potentially interesting, a different set of heuristics would be relevant in that setting.

Additionally, it is critical that nearly half of the lotteries were randomly generated for a prediction tournament because this guards against the concern that there is something peculiar about the lotteries that decision scientists invent and test, and indeed we know that lotteries are often designed to elicit particular responses, such as axiom violations. (I generate more lotteries in Section 6.3 to guard against this concern with respect to Independence.) While 171 lotteries may seem meager, pairing each lottery with every other lottery gives 29,070 choice pairs. This pairing method is another important safeguard against researcher-designed choices, because even when the original lottery pairs were designed to elicit a particular response, the present analysis ignores the intended pairings. It is true that offering every pair of lotteries results in some trivial choices—$1 for sure versus $1 million for sure—but these easy choices won't obscure behavior in more interesting cases, and indeed we will see that the PH in particular makes some surprising choices in cases that we might have considered uninteresting.

It is standard practice to study formalized lottery choice because lotteries capture the essential features of options, even if in many real-world situations those features can only be estimated. (Extrapolation of the results to more common situations is discussed in Section 7.) This paper is atypical in taking such a diverse set of lotteries and pairing each with each. It is more common to consider a restricted problem set in which all choices have a common feature (for instance, a fixed sum is compared to a risky lottery with similar EV); the purpose of this is to determine how well a given heuristic performs for different kinds of problems.

Since the purpose of this paper is different, it makes sense for the problem set to be different as well. The primary goal is to defend a methodological position about how heuristics should be compared, and not to characterize the circumstances in which any given heuristic should be used. To get results that hold broadly, the choice set must be correspondingly diverse. Perhaps the most significant division between lotteries is the magnitudes of the potential gains. I consider sub-contexts with respect to this by breaking the results down by choice EV in Section 5 and checking that payoff magnitude does not drive the results in Section 6.

Proponents of the PH in particular may object that the heuristic is meant to explain so-called 'hard' choices, meaning those in which the available lotteries have similar EVs. The important point here is that this paper is concerned with normative choice, and especially with determining the value of EU conformity once we grant that EV is relevant. From this viewpoint, the similar-EV choices that the PH best explains are unhelpful, precisely because the options are similarly good by design. Especially for the advocate of correspondence standards, the performance differences between heuristics will not reveal themselves on 'hard' choices. These choices are included in the test set, but they cannot comprise it.

4. EVALUATION BY THE EXPECTED UTILITY AXIOMS

4.1 Detecting Violations

As noted, two of the EU axioms—Transitivity and Independence—constrain the (rational) choice patterns of our heuristics. I, therefore, find all the violations of these two axioms for each heuristic, using a program created for the purpose. I discuss my methodological choices in more detail now. Sections 4.2 and 4.3 then present the results—the heuristics' axiom violation rates—for Transitivity and Independence, respectively.

The meaning and justification of Independence is taken to depend on its formulation, and concerns about the standard formulation have been raised which are relevant here. Specifically, there is the question of whether the lotteries in question are multi-stage lotteries or single-shot lotteries. Segal (1992) investigates the distinction thoroughly; in his terms, I use the "Mixture Independence Axiom" (171). Segal finds this version, which pertains to single-shot lotteries, to be less descriptively accurate and not as easily justified from a normative perspective. Nonetheless, the version I use is of interest because it is the standard version, and Section 6.3 provides the axiom with new support. Furthermore, the heuristics are only applicable to single-shot lotteries, so alternative formulations of Independence are not readily evaluated here.

Another methodological point is in order. I test the Independence and Transitivity axioms individually, rather than performing one test for compatibility with some EU hypothesis.[6] It is possible in principle for a choice pattern to violate EU Theory without explicitly violating any particular axiom; the violation may be implicit, as in "Zeckhauser's

[6] For those specifically interested in the PH, the author can provide recipes for holistic EU violations for lottery choices with up to three possible outcomes.

Paradox" (Jeffrey 1988).[7] Some axiom must be violated for EU to be violated, but the violation may be implicit. One might, therefore, worry that by testing Transitivity and Independence separately, I risk missing some violations.

Despite this, the results of this analysis are informative, and in some respects, the axiom-by-axiom method is preferable for present purposes. There are several reasons for this. Firstly, I reduce the risk of missing violations by directly testing for an important class of implicit violations, namely implicit Independence violations, as exemplified by the Allais Paradox.[8] Secondly, the sheer size of the heuristics' choice records means that each heuristic has ample opportunity to show its true colors in my tests. Thirdly, much of the discussion of EU's normative import focuses on particular axioms. For example, some people reject Independence, and Transitivity is often singled out in the debate between coherence and correspondence standards. The axiom-by-axiom analysis, therefore, facilitates engagement with the literature by revealing the performance of the heuristics with respect to particular axioms, and later, the costs of violating individual axioms.

4.2 Transitivity

This section shows the results of counting each heuristic's violations of Transitivity:

Transitivity If $A \succ B$ and $B \succ C$ then $A \succ C$

A set of choices violating Transitivity is a cycle. Cycles are impossible for every heuristic except for the PH, for the simple reason that each orders the lotteries according to a single number, such as the average of the 'probable' outcomes. Since it is lexicographic, the PH can violate Transitivity, and in fact, does so frequently: there are 101,253 total violations in the 876,044 cases of $A \succ B \succ C$ (so $C \succ A$ around 12% of the time).

[7] The Paradox is basically this: Suppose you are compelled to play Russian Roulette with a six-chamber revolver. Consider how much you would pay to remove the only bullet (guaranteeing your life), and how much you would pay to remove one bullet when five chambers are loaded. EU requires that you give the same price for each bullet, which contradicts typical price reports, but the violation must be derived.

[8] The Allais choices (see section 3.1) do not explicitly violate the axiom as written, but the violation can be quickly derived. Note that neither choice pair yields the other with some probability p; rather, the first pair has the form $(p \cdot A\ ;\ (1-p) \cdot C) \succ (p \cdot B\ ;\ (1-p) \cdot C)$ while the other has the form $(p \cdot A\ ;\ (1-p) \cdot D) \succ (p \cdot B\ ;\ (1-p) \cdot D)$. That $A \succ B$ is implicit.

This is already a striking finding because cycles are a cause for serious concern, but the source of these violations is also noteworthy; the PH produces many violations because it often makes choices that seem utterly unreasonable. Figure 1 shows a typical example (see the Appendix for more):

	Lottery
A	$10.60
B	$11.40 · (.97) ; $1.90 · (.03)
C	$310 · (.15) ; $230 · (.15); $170 · (.15); 130 · (.20) ; $0·(.35)

Figure 1 Lotteries between which the Priority Heuristic cycles

The PH chooses *A* over *B* because *A*'s minimum is nearly the maximum outcome for that pair, while *B*'s minimum is relatively small. Between *B* and *C*, the minima are similar and so they are not decisive. Instead, the heuristic checks the probability of those minima; the difference of .32 between the probabilities exceeds the threshold of .10, and so $B \succ C$ due to *B*'s smaller chance of earning the minimum. But the heuristic also chooses *C* over *A*: the minimum gains are not decisive, and *C* has a much more attractive probability of minimum gain (.35 instead of 1). Hence, the PH cycles.

This violation seems unrealistic, and the PH is billed as explanatory for choices between lotteries with similar EVs (Brandstätter et al. 2006, 24)—unlike many of the choices evaluated here. It bears repeating that I evaluate the heuristics in the abstract for a broad range of choice problems, and dissimilar-EV choices are normatively more interesting. I give practical conclusions for the PH in Section 7.2.

4.3 Independence

The test for Independence violations yields more mixed results. Recall the axiom:

Independence If $A \succ B$ then $(p \cdot A\ ;\ (1-p) \cdot C) \succ (p \cdot B\ ;\ (1-p) \cdot C)$

As noted above, two algorithms are used to detect both explicit and implicit violations. I only count strict Independence violations: it is not counted as a violation if $A \succ B$ and $A' \sim B'$, even if $A \succ B$ implies (via Independence) that $A' \succ B'$. This is a reasonable way to proceed because, first, a strict preference for the "wrong" lottery ($B' \succ A'$) is plausibly serious in a way that indifference between them is not. Moreover, EU states requirements on choices, and when a heuristic is indifferent a person applying it would choose one of the lotteries based on additional considerations; this choice simply goes beyond what the heuristic determines.

Minimax, Maximax, Most Likely, and Probable never produce strict Independence violations; the other heuristics do. The results of the evaluation are summarized in Figure 2, for both axioms. For reference, there are twelve unique opportunities to violate Independence.

Heuristic	**Transitivity**	**Independence**
Priority Heuristic	101,253	4
Minimax	0	0
Maximax	0	0
Hurwicz, α=.1	0	1
Hurwicz, α=.25	0	1
Hurwicz, α=.5	0	2
Hurwicz, α=.75	0	3
Hurwicz, α=.9	0	3
Equiprobable	0	4
Most Likely	0	0
Least Likely	0	7
Probable	0	0

Figure 2 Total axiom violations for each process. The maximum possible number of violations is 876,044 for Transitivity, and 12 for Independence.

On the one hand, some heuristics violate Independence at a high rate—for Least Likely, more than half the time. On the other hand, this rate could be misleading because the violations occur in cases designed to elicit them. Another reason for caution is the small sample: Independence places relatively few constraints on choices from the initial test set, and so the heuristics have relatively few violation opportunities. I avoid these problems when I measure the cost of Independence violations (see Section 6.3). Recall that Hurwicz takes the weighted average of the minimum and maximum outcomes, with weight α on the minimum. Figure 3 shows an example of a set of lotteries on which Hurwicz violates Independence for all the tested α values:

	Lottery
A	\$2500 · (.33) ; \$2400 · (.66) ; \$0 · (.01)
B	\$2400
C	\$2500 · (.33) ; \$0 ·(.67)
D	\$2400· (.34) ; \$0 · (.66)

Figure 3 Lotteries on which the Hurwicz Criterion violates Independence

For every α, *B*≻*A* (because 2400 is greater even than a .9 weighting of 2500) while *C*≻*D* (because 2500 is greater than 2400 no matter what their weighting). This choice pattern is an example of the 'Certainty Effect' (see Brandstätter et al. 2006, 11 for discussion). Independence requires, in contrast, that *A*≻*B if and only if C*≻*D*.

5. EVALUATION BY EXPECTED GAINS

5.1 Expected Value

I now compare the heuristics according to an objective wealth standard by comparing choice EVs. Specifically, given each pair of lotteries, the benchmark is the choice with the greater EV, along with the magnitude of that EV. When a heuristic is indifferent between two lotteries, we can simply average their EVs.

EV is an appropriate benchmark because it tells us the *expected* monetary value of each lottery, or equivalently, its cash equivalent for a risk-neutral individual, or its average payoff if it were played repeatedly. The greater the number of choices to be made, the less relevant variance becomes, and EV can be expected to coincide more closely with actual earnings. We need not simulate the lotteries themselves: by the Law of Large Numbers, the total profits for a heuristic's choices on the test set will be very close to the sum of the EVs of the chosen lotteries. This makes EV especially apt for processes that will be used repeatedly. Moreover, as an objective standard, there is no better option: accounting for an individual's attitude towards the variance would mean looking at their subjective preferences.

We can compare both the aggregate EVs (the total earnings each heuristic is expected to produce) and the average EVs (the percentage of the available EV that each heuristic realizes, averaged over all the choices). These quantities differ because in the aggregate case the relative impact of an individual choice depends on how much money is at stake, whereas in the average case each choice is equally important.[9] Nonetheless, the two measures support similar qualitative judgments, as Figures 4 and 5 demonstrate.

The presence of high-EV lotteries could distort our view because choosing a lottery with an EV of $500,000 when the other has an EV of $1 million leaves so much money on the table that it can overshadow performance on more modest (but perhaps more common) choices. The charts, therefore, break down the comparisons according to the maximum EV for each choice pair.

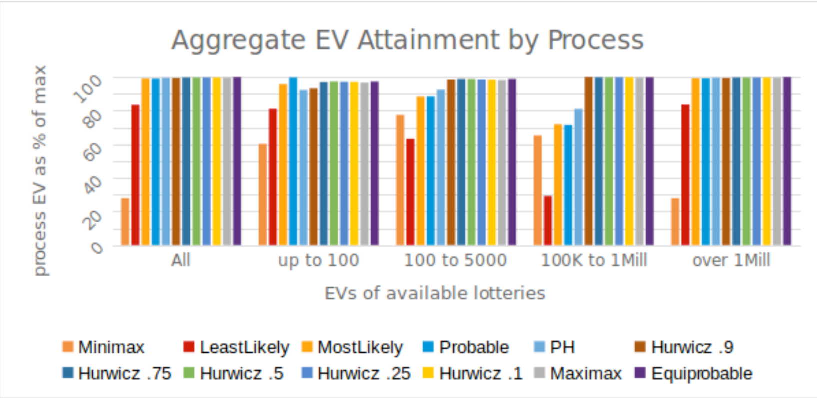


Figure 4 Aggregate EV attainment by process

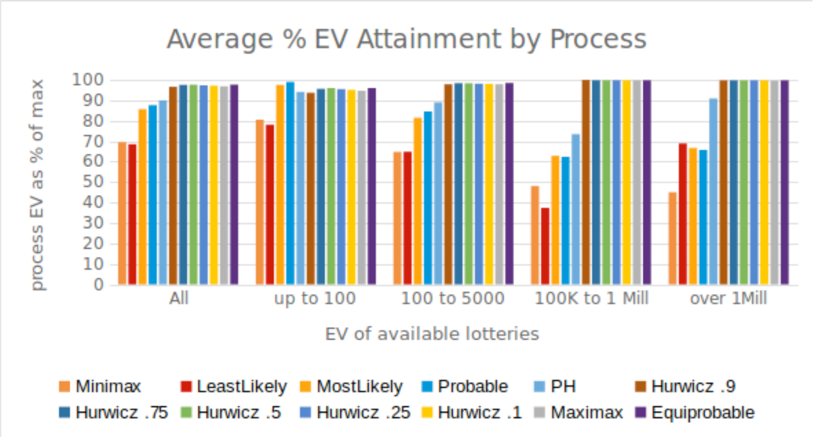


Figure 5 Average percentage EV attainment by process

These charts show that Minimax and Least Likely leave the most money on the table almost across the board; Minimax yields only 28% of

[9] For example, if a heuristic chooses a lottery with an EV of $3 million over one with an EV of $4 million, this registers as a $1 million loss in the aggregate and as a 25% loss as a percentage. In contrast, a choice of a $3 EV lottery over a $4 EV lottery has the same impact on the percentage performance, but a relatively tiny impact on the aggregate performance.

the total available EV, about 70% per choice on average, and leaves more as the lotteries get larger. This is to be expected, as these heuristics only consider the worst case and will avoid lucrative but riskier lotteries.

Most Likely, Probable, and the PH are overall the middle performers; they are noticeably better than Minimax and Least Likely, but noticeably worse than the rest. Most Likely and Probable are outperformed by Least Likely for the highest-valued choices, while Most Likely and Probable are the top performers for lotteries up to $100.

The overall top performers, then, are Equiprobable, Maximax, and the Hurwicz criteria, which all attain nearly maximal EV.[10] It is striking that all of them ignore probabilities entirely. Equiprobable does much better than Probable except in smaller lotteries; the only difference between them is that Probable ignores ‘improbable’ outcomes, but this leads it to ignore, for example, the 40% outcome in a 60/40 pair. Similarly, all instantiations of the Hurwicz Criterion do far better than Minimax, which shows that putting even a small amount of weight on the best possible outcome (as opposed to the worst) is sufficient to counteract Minimax’s caution and completely change the results.

Compared to the axiomatic performance metric, the major difference is that Minimax looks worst according to EV (whereas it was in perfect conformity with the EU axioms). Looking especially at Figure 5, the breakdown of EV attainment by EV bracket provides a partial explanation: as the maximum EV of the lotteries increases, Minimax becomes much less likely to choose the lottery with the higher EV. This is exactly what we would expect to see from a risk averse agent, for whom objective differences in EVs, especially high EVs, have much less subjective relevance than the differences between the minimum gains that can be guaranteed.

Comparing our two performance metrics, then, we see that it is possible to be perfectly coherent and yet not earn much (Minimax), perfectly coherent and a top earner (Maximax), and to violate occasionally and still be a top earner (Hurwicz). The least coherent heuristics—the PH and Least Likely—are only a moderate and a low earner, respectively. This suggests that incoherence is costly, but we need more data to prove this.

[10] This high performance is partly explained by the fact that many choices will involve simple dominance, and the heuristics will choose optimally in these cases. Such choices are ‘easy’ for all the heuristics, though, and will not affect the heuristics’ rankings.

5.2 Two Alternatives

One might not be entirely satisfied with the above evaluation. Firstly, one might think that Minimax is too unrealistic to be of much interest. Secondly, there is an alternative to EV that would better reflect typical preferences. This section briefly describes some supplementary analysis that addresses these points.

Although some people may be Minimax choosers, the vast majority are nowhere near so risk averse: this heuristic would choose a sure $1 over a lottery paying nothing or $1 million with equal probability. We can modify Minimax to create a more realistic conservative heuristic; call this TEV, EV combined with a threshold for the minimum gain. The intuition is this: an agent wants to ensure that their minimum gain meets a certain threshold—I use $1,000—so that they can pay a debt or take a trip, but after this aspiration is taken into account, they maximize EV. In my implementation, TEV chooses the lottery that guarantees at least $1,000 if only one does, and otherwise chooses according to EV. This hybrid process goes much of the way towards closing the gap between Minimax and the other processes because EV is often—but not always—the deciding factor; it yields 88% of the available EV in the aggregate, and 99% on average. So this sensible compromise between caution and EV maximization performs quite well.

As an alternative to EV, we could use a plausible utility function; the logarithmic utility function—proposed by Bernoulli (1954 [1738]), and frequently used in modern economics—is a natural candidate. It defines utility as $U(x)=ln(x)$, and since $ln(0)$ is undefined, I endow our hypothetical chooser with $10,000 in prior wealth to which their lottery earnings are added. The notable result is that Minimax performs much better, attaining approximately 98% of the available utility, because this utility function implies significant risk aversion. However, since this utility function is much more forgiving, it is also much less discriminating, and hence does not enable us to distinguish the heuristics very well by their performance.

6. REALIGNING COHERENCE AND CORRESPONDENCE

6.1 Overview

The axiom-based evaluation method and the correspondence evaluation method only partially agree about the ranking of the heuristics. What can be said about the precise relationship between these performance standards? From the coherence of a process, we can infer nothing about its EV attainment. The previous section raised the possibility that, nonetheless, incoherence might lead to diminished payoffs.

This section confirms that incoherence is costly by looking at the relationship between axiom violations and EV attainment for the most frequent axiom violators: the PH and Least Likely. Each choice that these heuristics make is taken as a data point, with two important attributes: first, is this choice associated with a violation (a binary variable); second, what percentage of the maximal EV does this choice attain in expectation? I evaluate Transitivity and Independence separately.

6.2 Transitivity

The correlation between cyclic choices and those that leave EV on the table tells a clear story. Only the PH is tested since it alone violates Transitivity. When the PH chooses $A \succ B$ and $B \succ C$, a choice of $C \succ A$ is associated with an EV loss of approximately 31%; the result is highly significant (p<.001) and the 95% confidence interval around the coefficient is quite tight. Controlling for other factors, such as minimum and maximum choice EV, does not change the result. (See the Appendix for more detail for both axioms.) Descriptive statistics tell the same story: when the PH violates Transitivity, the mean EV attainment of the choice is 64% of the maximum, while for non-violations the mean attainment is 95%. (Despite this significant finding, it is not the case that low-EV choices always coincide with a violation; the minimum attainment in each case is a small fraction of a percent. The point is that violations incur an EV loss, not the reverse.)

	Lottery	**EV**
A	\$10.60	\$10.60
B	\$11.40·(.97); \$1.90·(.03)	\$11.12
C	\$310·(.15); \$230·(.15); \$170·(.15); 130·(.20); \$0·(.35)	\$126

Figure 6 Priority Heuristic cycle with EVs

The statistical result also reflects what we observe when looking at violations. Recalling the example in Figure 6, the PH violates Transitivity relatively often because it is prone to making highly dubious choices; here, the $B \succ C$ choice is particularly costly. Again, the example is typical; the Appendix contains additional examples.

In fact, these statistics underestimate the cost of violating Transitivity because any given cycle involves three choices—hence three data points—but only one of those choices need be the 'mistake' responsible for the EV loss. This accounts for the fact that among both violating and non-violating choices, the median EV attainment is 100%. The 25th percentile of EV attainment among violating choices is only 8.5%, however, which fits perfectly with the hypothesis that the typical cycle contains two reasonable choices and one poor one.

6.3 Independence

The Independence violations from the initial task are too few for meaningful analysis. To remedy this, I construct new lotteries for which Independence has specific implications by compounding the lotteries in the initial set, following a (set of) patterns. Specifically, for each lottery A in the initial set, I create new lotteries $A' := pA\ ;\ (1-p)C$ for probabilities $p \in \{.1, .25\}$ and outcomes $C \in \{0, 25, 500, 5000, 5000000\}$. This means that each lottery A is associated with 10 additional lotteries A' to which it bears an Independence relationship, and every choice $A \succ B$ in the original set implies 10 additional choices $A' \succ B'$. The heuristics therefore have ample opportunity—up to 290,700 opportunities each—to violate Independence, providing enough data points to measure the cost of violations. This new data also provides a better test of Independence since the choice pairs were not designed to generate violations by human subjects.

Apart from this, the Independence analysis mirrors that for Transitivity. Over the 10 A' variations tested, the PH produces between 866 and 11,546 violations per variation (out of 29,070 choices for each). For Least Likely, the minimum is 106 and the maximum is 8,504 (note that Least Likely is more often indifferent between lotteries, so it has fewer opportunities for violations).

For both heuristics, violations are costly. A PH violation is associated with an EV cost of about 32%, and the result is highly significant ($p<.001$) with the 95% confidence interval narrowly around the coefficient. PH choices that violate Independence (that is, $B' \succ A'$ when $A \succ B$) yield only 66% of the available EV on average, while non-violations yield 99%. The median choice attains 93% of the available EV among violating choices, and 100% among non-violating choices. An Independence violation by Least Likely is associated with an EV cost of about 41%, again with $p<.001$ and a narrow confidence interval. Violating choices yield 46% of EV on average (and 41% at the median), while non-violating choices yield 87% (and a median of 100%). As with Transitivity, violations are the driver of cost even when EV is controlled for.

Although Independence violations can be famously compelling (as with the Allais paradox), the reason for their cost is straightforward. Suppose a lottery *A* has greater EV than another *B*, and in fact *A* is chosen. Now, for any *p* and *C* used to make compounds *A*' and *B*', *A*' will have a higher EV than *B*'. An Independence violation, therefore, *guarantees* that one choice fails to maximize EV; the same is true for Transitivity violations. While it is prima facie legitimate to question money pump and Dutch Book arguments, these arguments are essentially elaborations of this observation. The significance of the results presented here lies in the magnitude of the cost, not its mere existence.

7. DISCUSSION

7.1 How to Evaluate Heuristics

This case study demonstrates that the hybrid approach combining EU and ER avoids their individual problems and is more informative regarding the performance of heuristics. Aside from making a small concession to ER by evaluating heuristics, an otherwise pure EU approach would simply rank the heuristics according to how often they generate incoherent (EU-violating) choices; thus, perfectly coherent heuristics would be deemed perfectly rational. In contrast, a pure ER approach would rank the heuristics strictly according to objective performance criteria such as their EV attainment. Additional rankings might be produced to account for additional virtues such as speed. These rankings would similarly be based on objective measurements, such as the average number of computational steps. (Here, the heuristics are all so fast that speed is essentially irrelevant.) The most ecologically rational heuristics would achieve the best EV/speed combination.

Both pure approaches yield evaluations with critical flaws. For example, EU judges both Maximax and Minimax to be perfectly rational. These heuristics imply very different preferences, though, and so each will be wrong for many people. For some, Minimax would guarantee inadequate earnings, while for others Maximax would involve an unacceptable risk of the same fate. ER judges heuristics more favorably the more closely they coincide with EV maximization. Again, this is wrong for agents who are not risk-neutral, for example those for whom

$2 million and $4 million have practically equal utility. For them, a conservative heuristic would be more appropriate.

By combining the EU and ER tests, we achieve a fuller picture of the heuristics' performance and can avoid both kinds of mistakes. Furthermore, by assessing the cost of incoherence according to the correspondence standard, we can determine how relevant the EU standard is even for those who prioritize objective success. The best implementation of the hybrid approach is therefore to perform all three tests (when possible), thereby extracting all of the potentially valuable information from the choice data so that theorists and agents can make informed decisions about which heuristics to use, endorse, and teach.

Yet there is another equally important aspect of the hybrid approach, and an equally important lesson to be learned from this case study. In the case of lottery choices, both coherence and correspondence standards are readily available. For many problems of interest, however—and especially for the kinds of real-world problems of interest to ER proponents—correspondence standards are harder to come by. Coherence can then serve as a proxy—just as Hammond (1996, 2007) argues—bolstered by the demonstrated connection between incoherence and diminished performance. (Of course, the strength of this connection may vary with context, which is why the connection itself should be tested whenever possible.)

As an example, many real-life decision problems involve not risk but rather uncertainty, where the probabilities of the possible outcomes are not known and can only be estimated with more or less confidence. Even if agents have valid subjective probabilities, these are unknown to (and hence unusable by) the theorist in comparing the heuristics, just as with subjective utility. Such decision problems are less amenable to simulation and objective ranking. Nonetheless, the coherence test provides a way to compare possible heuristics, and the connection between coherence and correspondence in the case of risk—especially since the connection is demonstrated for a very broad context—provides evidence that less coherent heuristics would yield objectively worse results in the case of uncertainty as well. Let us now turn to the relevance of the hybrid approach for people making real-world choices.

7.2 How To (Help People) Choose

These simulations do not permit fine distinctions regarding heuristic performance in specific contexts. Further studies would do so, but no study would determine the 'best' heuristic, even relative to a context. Instead, agents must choose heuristics that align with their expectations, preferences, and aspirations in their particular choice context. This fits well with the motivation expressed by Thorngate (1980), Johnson and Payne (1985), and Bordley (1985)—and the intervening years have seen growing interest in helping the public in this way.

Let us first consider how we should respond to the PH's performance. The PH performs poorly here, but it is hypothesized that people use it for similar-EV and therefore less critical choices. This illustrates an important point, which is that the first step of any attempt to improve people's choices should be to determine whether their current choices are especially problematic. It is only worth investing limited time and resources to teach people new heuristics in cases where their existing choice processes are likely to serve them especially poorly. Absent evidence that people are often unhappy with the outcomes of their PH choices, we should not see the use of this heuristic as especially problematic.

Some of the heuristics assessed here—namely those from traditional decision theory—have long been evaluated on their theoretical merits and through examples and intuition, but the simulation method allows us to assess them according to how well we can actually expect them to perform. The results indicate that simpler is often better, Maximax being the most extreme example; this is convenient because simpler is also easier to learn. In contrast, the PH underperforms in an important sense because it is more complicated: its lexicographic nature enables it to make costly intransitive choices.

At this point, one might ask why agents should not simply be taught to maximize EV, at least as a first step. Estimating and attempting to maximize EU is probably not feasible for people without significant formal training, but the arithmetic required to calculate EV is simple, and the exercise would provide a valuable safeguard against very bad choices. EV also basically dominates Equiprobable. While people ought to learn the basics of EV calculation in school along with some fundamentals of probability, there are broader advantages to learning simple heuristics too.

Again, an important consideration is that uncertainty is more common than risk in everyday life, but EV cannot be calculated under uncertainty. Subjective probabilities may be inaccurate, incoherent, or inaccessible. Heuristics become more attractive as choices get more complicated in this way. Since many of the heuristics studied here perform well given a range of probability and outcome distributions, and the top performers make no use of probabilities anyhow, we can extrapolate from their performance here and expect those heuristics to do well in situations of true uncertainty.

Considering situations of uncertainty makes Minimax look even more appealing. In situations of risk, this heuristic epitomizes the tension between coherence and correspondence. In situations of radical uncertainty, however, EV is irrelevant (note that this also makes Equiprobable more reasonable). For an agent whose priority is to make conservative choices, Minimax could be an excellent choice: it involves practically no effort, it minimizes risk, and it will not lead the agent into incoherence (which would entail a cost). Minimax could also easily be used conditionally—as in TEV—by an agent who is risk averse only below a certain aspiration level, or when losses are possible. For those seeking a less conservative heuristic, the Hurwicz Criterion could be very useful. It allows the agent to choose exactly how much weight to put on the worst outcome, and how much on the best; this balance could even be varied contextually. While this heuristic is not perfectly coherent, it can accommodate conservative preferences to a high degree and promises much higher earnings than Minimax. By evaluating these heuristics with the hybrid approach, we are in the best position to help choosers to find their preferred balance.

APPENDIX: VIOLATIONS AND THEIR COSTS

Transitivity

Additional examples of Transitivity violations by the Priority Heuristic:

A	\$15.50
B	\$18.90 · .9 ; \$6.70 · .1
C	\$1000 · .5 ; \$0 · .5

A	\$3,000 · .002 ; \$0 · .998
B	\$10.60
C	\$17.90 · .92 ; \$7.20 · .08

A	\$15.50
B	\$18.90 · .9 ; \$6.70 · .1
C	\$5000000 · .1 ; \$0 · .9

Figure 7 Priority Heuristic cycles

The correlations described in Section 6.2 are based on the following table:

Call:				
glm (formula = PhexpPrcntgofMax ~ PHtransViolYN)				
Deviance Residuals:				
Min	1Q	Median	3Q	Max
-94.868	5.132	5.132	5.132	35.814
Coefficients:				
	Estimate	Std. Error	t value	Pr (> | t |)
(Intercept)	94.86825	0.02756	3441.9	<2e – 16 ***
PHtransViolYN	-30.68178	0.08103	-378.6	<2e – 16 ***

Figure 8 Priority Heuristic Transitivity regression table

Due to the very large sample size, use the standard z^* for the 95% confidence interval as the critical value. Let β^* be the correlation coefficient and *se* the standard error. Then the above yields $\beta \in [\,\beta^* \pm z^* \cdot \text{se}] = [-30.68 \pm 1.96 \cdot .08] = [-30.83\ ,\ -30.52]$ as the 95% confidence interval around the regression coefficient of -30.68 (in other words, a violation is associated with a decrease of 30.68% EV).

Independence

The correlations reported in Section 6.3 are based on the regression shown in Figures 9 and 10.

Again use the standard z* for the 95% confidence interval as the critical value. Let β^* be the correlation coefficient and *se* the standard error. Then the above yields $\beta \in [\,\beta^* \pm z^* \cdot \text{se}\,] = [\,-31.94 \pm 1.96 \cdot .09\,] = [\,-32.12\ ,\ -31.76]$ as the 95% confidence interval around the regression coefficient of -31.94 (i.e. a violation is associated with a decrease of 31.9% EV), for the PH. For Least Likely, the calculation is $\beta \in [\,\beta^* \pm z^* \cdot se] = [\,-40.74 \pm 1.96 \cdot .18\,] = [\,-40.09\ ,\ -40.38\,]$.

Call:				
lm (formula = PrcntEV ~ ViolCode)				
Residuals:				
Min	1Q	Median	3Q	Max
-98.503	1.497	1.497	1.497	33.439
Coefficients:				
	Estimate	Std. Error	t value	Pr (> | t |)
(Intercept)	98.50283	0.04148	2374.6	<2e – 16 ***
ViolCode	-31.94203	0.09345	-341.8	<2e – 16 ***

Figure 9 Priority Heuristic Independence regression table

Call:				
lm (formula = LL_PrcntEV ~ ViolCode)				
Residuals:				
Min	1Q	Median	3Q	Max
-86.578	6.622	13.422	13.422	54.162
Coefficients:				
	Estimate	Std. Error	t value	Pr (> | t |)
(Intercept)	86.57767	0.06247	1386	<2e – 16 ***
ViolCode	-40.74004	0.17951	-227	<2e – 16 ***

Figure 10 Least Likely Independence regression table

REFERENCES

Allais, Maurice. 1953. "Le Comportement de L'Homme Rationnel Devant Le Risque: Critique Des Postulates et Axiomes de L'Ecole Americaine." *Econometrica* 21 (4): 503–46.

Arkes, Hal R., Gerd Gigerenzer, and Ralph Hertwig. 2016. "How Bad Is Incoherence?" *Decision* 3 (1): 20–39.

Berg, Nathan. 2014. "The Consistency and Ecological Rationality Approaches to Normative Bounded Rationality." *Journal of Economic Methodology* 21 (4): 375–95.

Berg, Nathan, Guido Biele, and Gerd Gigerenzer. 2016. “Consistent Bayesians Are No More Accurate Than Non-Bayesians: Economists Surveyed About PSA.” *Review of Behavioral Economics* 3 (2): 189–219.

Berg, Nathan, and Gerd Gigerenzer. 2006. “Peacemaking Among Inconsistent Rationalities?” In *Is There Value in Inconsistency?*, edited by C. Engel. and L. Daston. Baden-Baden: Nomos.

Bernoulli, Daniel. “Exposition of a New Theory on the Measurement of Risk.” *Econometrica* 22 (1): 23–36.

Binmore, Ken. 2009. *Rational Decisions*. Princeton: Princeton University Press.

Bordley, Robert F. 1985. “Systems Simulation Comparing Different Decision Rules.” *Behavioral Science* 30 (4): 230–39.

Brandstätter, Eduard, Gerd Gigerenzer, and Ralph Hertwig. 2006. “The Priority Heuristic: Making Choices Without Trade-Offs.” *Psychological Review* 113 (2): 409–32.

Fishburn, Peter C. 1989. “Retrospective on the Utility Theory of Von Neumann and Morgenstern.” *Journal of Risk and Uncertainty* 2 (2): 127–158.

Gigerenzer, Gerd, Ralph Hertwig, and Thorsten Pachur 2011. *Heuristics: The Foundations of Adaptive Behavior*. New York: Oxford University Press.

Gigerenzer, Gerd, and Reinhard Selten. 1999. *Bounded Rationality: The Adaptive Toolbox*. Cambridge: The MIT Press.

Gilboa, Itzhak. 2009. *Theory of Decision Under Uncertainty*. Cambridge: Cambridge University Press.

Hammond, Kenneth R. 1996. *Human Judgment and Social Policy: Irreducible Uncertainty, Inevitable Error, Unavoidable Injustice*. Oxford: Oxford University Press.

Hammond, Kenneth R.. 2007. *Beyond Rationality: The Search for Wisdom in a Troubled Time*. Oxford: Oxford University Press.

Hastie, Reid, and Kenneth A. Rasinski. 1988. “The Concept of Accuracy in Social Judgment.” In *The Social Psychology of Knowledge*, edited by Daniel Bar-Tal and Arie W. Kruglanksi, 193-208. New York: Cambridge University Press.

Hájek, Alan. 2009. “Dutch Book Arguments.” In *The Handbook of Rational and Social Choice*, edited by Paul Anand, Prasanta Pattanaik, and Clemens Puppe, 173–96. Oxford: Oxford University Press.

Jeffrey, Richard. 1988. “Biting the Bayesian Bullet: Zeckhauser’s Problem.” *Theory and Decision* 25: 117–22.

Johnson, Eric J., and John W. Payne. 1985. “Effort and Accuracy in Choice.” *Management Science* 31 (4): 395–414.

Kahneman, Daniel, and Amos Tversky. 1979. “Prospect Theory: An Analysis of Decision Under Risk.” *Econometrica* 47 (2): 263–92.

Luce, R. Duncan, and Howard Raiffa. 1957. *Games and Decisions: A Critical Survey*. Mineola: Dover Publications.

Markowitz, Harry. 1952. “The Utility of Wealth.” *Journal of Political Economy* 60 (2): 151–58.

Mas-Colell, Andreu, Michael D. Whinston, and Jerry R. Green. 1995. *Microeconomic Theory*. New York: Oxford University Press.

Neumann, John von, and Oskar Morgenstern. 1944. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Rich, Patricia. 2016. "Axiomatic and Ecological Rationality: Choosing Costs and Benefits." *Erasmus Journal for Philosophy and Economics* 9 (2): 1–33.

Rich, Patricia. 2018. "Comparing the Axiomatic and Ecological Approaches to Rationality: Fundamental Agreement Theorems in SCOP." *Synthese* 195 (2): 529–47.

Segal, Uzi. 1992. "The Independence Axiom Versus the Reduction Axiom: Must We Have Both?" In *Utility Theories: Measurements and Applications*, edited by Ward Edwards, 165–83. Boston: Kluwer Academic Publishers.

Sturm, Thomas. 2012. "The 'Rationality Wars' in Psychology: Where They Are and Where They Could Go." *Inquiry* 55 (1): 66–81.

The Technion Prediction Tournament, 2008. Organized by Ido Erev, Eyal Ert and Alvin E. Roth. Accessed June 2015. Retrieved from <tx.technion.ac.il/erev/COMP/>.

Thorngate, Warren. 1980. "Efficient Decision Heuristics." *Behavioral Science* 25 (3): 219-25.

Tversky, Amos. 1972. "Elimination by Aspects: A Theory of Choice." *Psychological Review* 79 (4): 281–99.

Tversky, Amos, and Daniel Kahneman. 1992. "Advances in Prospect Theory: Cumulative Representation of Uncertainty." *Journal of Risk and Uncertainty* 5: 297–323.

Von Neumann, John, and Oskar Morgenstern. 1944. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Wallin, Annika. 2013. "A Peace Treaty for the Rationality Wars? External Validity and Its Relation to Normative and Descriptive Theories of Rationality." *Theory & Psychology* 23 (4): 458-78.

Patricia Rich is a researcher on the DFG project *Knowledge and Decision* (P.I. Moritz Schulz) in the Philosophy Department at the University of Hamburg. Her main research areas are epistemology and game and decision theory, with a focus on interdisciplinary methodology. In addition to the present topic, she has written on belief revision in games and reputation in signaling games.
Contact e-mail: <patricia.rich@uni-hamburg.de>

Erasmus Journal for Philosophy and Economics,
Volume 11, Issue 1,
Spring 2018, pp. 31-49.
https://doi.org/10.23941/ejpe.v11i1.285

The Evolutionary Explanation of What? A Closer Look at Adaptationist Explanations of Risk Preferences

BENGT AUTZEN
Munich Center for Mathematical Philosophy

Abstract: The paper examines evolutionary explanations of risk preferences. First, the paper argues that evolutionary psychology is ill-suited for explaining prospect theory risk preferences since the empirical evidence does not support the universality of the fourfold pattern of risk preferences postulated by prospect theory. Second, the paper argues that explaining prospect theory risk preferences by means of risk-sensitive foraging models is incomplete since this approach does not offer a rationale for the observed diversity in human decision making involving monetary gambles. Finally, the paper suggests adopting a wider perspective on evolutionary approaches to human behaviour that also takes into account the role of cultural processes in shaping risk preferences.

Keywords: risk preferences, prospect theory, evolutionary psychology, human behavioural ecology, cultural evolution.

JEL Classification: D01, D91, Z10

I. INTRODUCTION

In common parlance risk refers to the possibility of harm, injury or loss. Among decision theorists and economists, however, risk is associated with a different concept. Rather than identifying risk with the possibility of harm, risk refers to uncertainty or, more precisely, the dispersion of outcomes in a probability distribution. As such, risk is typically associated with statistical concepts such as the variance of a probability distribution. Formally, an agent is said to be risk-averse if and only if

Author's Note: I would like to thank Alasdair Houston, Samir Okasha, Patricia Rich and Armin Schulz as well as the editors and anonymous referees of the journal for helpful comments on earlier versions of the manuscript. Awards from the British Academy Postdoctoral Fellowship Scheme and the Alexander von Humboldt Foundation are gratefully acknowledged.

she prefers x for certain to a lottery with expected monetary value x. An agent is said to be risk seeking if and only if she prefers a lottery with expected monetary value x to x for certain. While it is rather intuitive that human beings are averse to risk when interpreted as the possibility of harm, it is an open question of whether—and if so, why—human beings are averse to risk when interpreted as the dispersion of outcomes.[1]

Okasha (2007) offers an adaptationist explanation of risk aversion that invokes results from theoretical biology; these results demonstrate that natural selection is sensitive to both the mean and the variance of the offspring distribution when organisms evolve in stochastic environments. In particular, given two traits with the same mean offspring number, it can be shown that under certain environmental conditions natural selection favours the trait with the lower variance in reproductive success. Okasha's account has been criticised on the grounds that it misconstrues its explanandum. Rather than explaining that human beings are risk-averse, Schulz (2008) argues that explaining human attitudes towards risk requires the explanation of both risk-averse and risk-seeking behaviours.

Prospect theory is generally considered to be the most influential descriptive account of decision making under risk in psychology and behavioural economics (Tversky and Kahneman 1992). The theory stipulates that for events with moderate to high probability agents are risk-averse in the gains domain and risk-seeking in the loss domain. For gains and losses with low probability, however, the pattern is reversed. This postulated attitude towards risk is referred to as the fourfold pattern of risk preferences.[2]

In line with Schulz's requirement that an adequate evolutionary explanation of risk preferences has to account for both risk-averse and risk-seeking behaviour in human agents, a number of evolutionary explanations of the risk preferences postulated by prospect theory have been proposed (Aktipis and Kurzban 2004; Brennan and Lo 2011; McDermott et al. 2008; Mishra and Fiddick 2012; Mallpress et al. 2015). Mallpress et al. (2015), for instance, provide an adaptive rationale for the fourfold pattern of risk preferences by identifying conditions under

[1] For a more detailed treatment of these two different concepts of risk (i.e., risk as the possibility of harm and risk as dispersion), see Friedman et al. (2014).

[2] More precisely, the focus here is on what is referred to as cumulative prospect theory (Tversky and Kahneman 1992).

which these risk preferences maximize the reproductive value of a decision maker. Mallpress et al. demonstrate that prospect theory risk preferences can arise when environmental conditions change stochastically over time, thereby affecting the reserve energy level of a decision maker, and the pattern of change shows auto-correlation.[3]

In order to further the philosophical debate on the evolution of human attitudes towards risk, I will take a closer look at evolutionary explanations of prospect theory risk preferences. I will make three points. First, I will argue that evolutionary psychology is ill-suited for explaining prospect theory risk preferences since the empirical evidence does not support the universality of the fourfold pattern of risk preferences. Second, I will argue that explaining prospect theory risk preferences by means of risk-sensitive foraging models is incomplete since this approach does not offer a rationale for the observed diversity in human decision making involving monetary gambles. And third, I will suggest adopting a wider perspective on evolutionary approaches to human behaviour that also takes into account the role of cultural processes in shaping risk preferences.

The structure of the paper is as follows. Section 2 introduces basic ideas from evolutionary psychology as well as some criticisms raised against this evolutionary approach to human behaviour. Section 3 revisits the evidence for the risk preferences postulated by prospect theory. Section 4 turns to the application of human behavioural ecology to the study of prospect theory risk preferences. Section 5 offers some suggestive remarks on what the literature on cultural evolution can contribute to our understanding of human attitudes towards risk. Section 6 concludes with some general thoughts on the prospect of explaining risk preferences evolutionarily.

II. EVOLUTIONARY PSYCHOLOGY AND RISK PREFERENCES

Evolutionary psychology studies how organisms adapt behaviourally to their environment. The focus of evolutionary psychology has traditionally been on universal adaptations, that is, aspects of the human genome that became fixated in the population by natural

[3] In a discrete model environmental states are said to be positively auto-correlated if the occurrence of a given environmental state at time t increases the probability of the same state occurring at time t + 1. In such a setting knowledge about a current environmental state provides information about the likely environmental conditions in the near future.

selection before our species spread across the world about 50,000 years ago and that have not changed systematically since (Tooby and Cosmides 1990). For instance, evolutionary psychologists have proposed two central hypotheses regarding sex differences in human mating preferences. It has been argued that men have an evolved preference for mating with young women while women have an evolved preference for mating with high-status men (e.g., Buss 1992; Ellis 1992).

Aktipis and Kurzban (2004) suggest that evolutionary psychology can provide evolutionary explanations of human preferences, including human attitudes towards risk. They write:

> Economists can (and do) claim that individuals get utility from these activities, leaving the question of the origin of tastes and preferences to the other behavioral sciences [...]. Evolutionary psychology provides answers—or at least a way to generate possible answers—about these origins tastes and preferences that enabled us to better solve adaptive problems were selected for during human evolutionary history (Aktipis and Kurzban 2004, 137).

Similarly, McDermott et al. (2008) motivate their evolutionary account of prospect theory preferences by reference to work in evolutionary psychology. More specifically, they argue that the human cognitive architecture evolved to solve particular adaptive problems related to finding sufficient food resources required for survival persists and is currently utilized in other survival-related decisions.

In order to establish my critique of the use of evolutionary psychology for the explanation of prospect theory risk preferences, I will begin with a criticism of evolutionary psychology originally due to Buller (2005). Buller questions whether the mating preferences postulated and subsequently explained by evolutionary psychologists constitute a trait that is universally shared by human beings. He develops his objection by first setting a standard that mating preferences have to satisfy in order to be considered as universal. Buller writes:

> to say that those preferences are "universal" means that they are observable *in all cultures, all historical periods, all economic or political systems, all social classes, all religious groups, all "races" or ethnicities, and all relevant ages of the life cycle* (2005, 210, italics in original).

In a second step, Buller argues that evolutionary psychologists have failed to provide evidence that the mating preferences inferred by evolutionary psychologists are universal among humans given this standard. In particular, he argues that mating preferences tend to vary with age and social class. As such, evolutionary psychologists have misconstrued the explanandum of their account of mating preferences.[4]

From the perspective of this paper, the second step of Buller's objection is of less relevance since it is specific to the subject of human mating preferences. The first step, however, raises a more general issue that also applies to evolutionary explanations of human attitudes towards risk based on ideas from evolutionary psychology. In order to assess whether prospect theory risk preferences fall into the category of universal preferences that shape human nature, a notion of universality has to be adopted that allows for empirical data to have a say on the subject matter. Obviously, such an account of universality should not be overly restrictive in order to provide a convincing critique of evolutionary psychology. Here, I will adopt the requirement that our best available evidence has to support the idea that a majority of agents adopts the fourfold pattern of risk preferences when making decisions under risk.

The insistence that a particular preference is shared by the majority of agents introduces some arbitrariness into the discussion. In particular, one might wonder what makes the 50% threshold philosophically relevant. This requirement, however, sits well with recent characterizations of the subject matter of evolutionary psychology arguing that evolutionary psychology aims to explain traits that are present in most humans (Machery 2008). Based on Machery's conception, the focus of evolutionary psychology is on the similarities between humans rather than on human differences. This step does not deny that evolutionary biology can explain polymorphisms found in the human population, such as the differences in blood types. Furthermore, this view does not ignore that some evolutionary psychologists have recently turned to providing selective accounts for individual

[4] Buller's critique has faced a number of objections. For instance, Delton et al. (2006) argue that Buller wrongly assumes that the mating preferences postulated by evolutionary psychology are to be identical across different stages of the life cycle. That being said, Buller's argument offers a good starting point for assessing the application of evolutionary psychology to the explanation of human attitudes towards risk.

differences (Buss and Hawley 2010) but delegates the explanation of these phenomena to other evolutionary approaches.

My reading of the notion of universal preference is also rather modest in a different sense. I do not ascribe a particular reading to the notion of preference but only tacitly assume a notion of preference that is compatible with stochastic choice models in economics.[5] Doing so, however, requires a notion of preference that allows for the possibility that agents can make errors in their choice behaviour. That is, even though an agent can have preferences satisfying various axioms of rational choice theory, she can wrongly express these preferences in a choice situation. While this restriction does not impact the course of this paper, it is worth noting that this constraint rules out revealed preference theory as a matter of logic. According to revealed preference theory, preference is reducible to (hypothetical) choice. That is, a preference ordering over a set of alternatives is just a summary of an agent's choices between them. One consequence of this reading is that agents cannot, by definition, make mistakes when expressing their true preferences.

III. EVIDENCE AND PROSPECT THEORY

A comprehensive review of the literature and a recent experimental study on the evidential basis of prospect theory has been provided by Harrison and Swarthout (2016). My presentation of Harrison and Swarthout's work follows the summary of Harrison and Ross (2017). Harrison and Swarthout argue that virtually no previous studies have estimated a model of prospect theory in which all experimental tasks involved real payoffs, and that those studies that were satisfactory from a methodological perspective found little evidence in support of the theory. Based on their experimental data, Harrison and Swarthout conclude that human decision making under risk is heterogeneous and almost all of the experimental subjects apply rank-dependent utility theory rather than prospect theory or, to a lesser extent, expected utility theory rather than prospect theory. Rank-dependent utility theory proposed by Quiggin (1982) extends orthodox expected utility theory by allowing for decision weights on lottery outcomes. As such, rank-dependent utility theory transforms probabilities into decision weights similar to prospect theory. In contrast to prospect theory, however,

[5] A number of stochastic choice models have been proposed in the decision-theoretic literature. For an overview, see Suppes et al. (1989) and Wilcox (2008).

rank-dependent utility theory does not invoke the concept of a reference point based on which gains and losses are to be evaluated. Similar to prospect theory, rank-dependent utility theory is designed to make sense of the fact that agents both purchase lottery tickets and insure against losses. Harrison and Swarthout's data suggest that most of the apparently loss-averse choice behaviour results from probability weighting rather than from direct disutility experienced when an outcome is framed against a reference point. That is, their experimental subjects behave as if they evaluate the net payment rather than the gross loss when one is presented to them and then apply probability weighting consistent with rank-dependent utility theory.

Prospect theory is widely seen as the most promising descriptive account of decision making under risk. In the light of the existing literature as well as some recent experimental work, however, the laboratory evidence is not as solid as previously assumed. According to Harrison and Swarthout's experimental study, the most empirically adequate hypothesis about human choice under risk is that it is heterogeneous and that in cases where agents do not follow expected utility theory, choice behaviour is better characterized by rank-dependent utility theory than prospect theory. I conclude that a majority of agents in these experiments appear to follow decision making models different from prospect theory. As such, the empirical evidence does not support the idea that humans universally share the fourfold pattern of risk preferences. This suggests that evolutionary psychology—understood as an evolutionary account of universal human traits—is not the right theoretical framework to produce an evolutionary explanation of the fourfold pattern of risk preferences.

This conclusion sounds familiar from the perspective of earlier philosophical critiques of evolutionary approaches to human behaviour. In the context of sociobiology, Gould and Lewontin (1979) as well as Kitcher (1985) have identified a flawed form of scientific reasoning that combines an overly liberal form of evolutionary thinking with loose experimental testing. More specifically, they argue that sociobiologists have been culpable of providing spurious confirmation to the existence of traits whose empirical basis is rather weak. While I have not taken issue with evolutionary models giving rise to prospect theory preferences, such as Mallpress et al. (2015), I have also diagnosed a lack of empirical support for the preferences explained by these models.

One might object to my critique that observing choice behaviour, which seems to follow diverse models of decision making under risk, is not sufficient to rule out the universality of prospect theory risk preferences. Indeed, Cosmides and Tooby (1997) stress that cultural diversity is compatible with the existence of a set of universal cognitive adaptations. For instance, they argue that humans share a preference for sweet foods but that the expression of this preference has changed significantly since the Pleistocene. Modern humans have a large number of different food options compared to their hunter-gatherer ancestors and as a result their preference for sweet foods manifests itself in different ways (e.g., in the consumption of fast food). Returning to the evolution of risk preferences, however, it is unclear whether a similar argument can be made. Taking the analogy with the universality of the preference for sweet food seriously, would require that the fourfold pattern of risk preference constitutes the universal human attitude towards risk while choice behaviour that is more aptly characterized as following expected utility theory or rank-dependent utility theory corresponds to the different manifestations of this preference in contemporary society. Since it is difficult to make sense of risk preferences that follow a rank-dependent utility model as a manifestation of the risk preferences postulated by prospect theory, the analogy between the universality of food and risk preferences breaks down.

IV. RISK-SENSITIVE FORAGING MEETS PROSPECT THEORY

Evolutionary psychology traditionally focuses on universal features of human psychology in its explanations. In contrast to evolutionary psychology, human behavioural ecology aims to provide adaptationist accounts of the observed differences in human behaviour. Laland and Brown write:

> The principal goal of human behavioural ecology is to account for the variation in human behaviour by asking whether models of optimality and fitness-maximisation provide good explanations for the differences found between individuals. An overriding assumption is that human beings exhibit an extraordinary flexibility of behaviour, allowing them to behave in an adaptive manner in all kinds of environments (2002, 112).

Similarly, Smith et al. characterise the explanatory strategy employed by human behavioural ecologists as follows:

> [Human behavioural ecology] applies the theoretical perspective of animal behavioral ecology to human populations, examining the degree to which behavior is adaptively adjusted to environmental (including social) conditions, emphasizing conditional strategies of the form "in situation X, maximize fitness payoffs by doing α, in situation Y, do β" (2001, 128).

Evolutionary psychologists stress that the environment of contemporary human beings differs substantively from the selective environment faced by our ancestors, which is typically understood as the Pleistocene environment inhabited by our hunter-gatherer ancestors. As a result, evolutionary psychologists postulate an adaptive lag between the environment during which complex human behavioural traits have been shaped by natural selection and the present-day environment inhabited by modern human beings. Human behavioural ecologists, on the other hand, downplay the significance of this adaptive lag. From their perspective, evolutionary psychologists underestimate the amount of currently adaptive behaviour found in the human population.[6]

Human behavioural ecologists regularly employ risk-sensitive foraging theory in their models. Risk-sensitive foraging theory provides an account of how animals should choose between stochastic foraging options in order to maximize reproductive success (Caraco 1980; Stephens 1981; McNamara and Houston 1992). A particular risk-sensitive foraging model that has been invoked in evolutionary explanations of prospect theory risk preferences is the daily energy budget rule due to Stephens (1981). This decision rule aims to explain the behaviour of small birds foraging during the winter months.

The problem faced by these birds is that they need to acquire enough energy during the day in order to survive the following night. Suppose the foraging bird has two foraging options that have the same expected energy gain but differ in variance. Stephens shows that the foraging bird should choose the more variable foraging option if the daily energy budget is negative, that is, if the expected energy gains are insufficient to meet the energy requirements, and the less variable option if the daily energy budget is positive.

A number of researchers in the social sciences have made use of results from risk-sensitive foraging theory in order to offer an

[6] For a more comprehensive treatment of the differences between evolutionary psychology and human behavioural ecology, see Laland and Brown (2002).

evolutionary rationale for prospect theory. For instance, Aktipis and Kurzban (2004) argue that the asymmetry between losses and gains postulated by prospect theory is underwritten by risk-sensitive foraging theory since marginal energy losses are more fitness relevant than marginal fitness gains. While energy losses can sometimes lead to death, energy gains will merely extend the life span of a forager. Furthermore, they assert that the curvature of the value function in prospect theory is underwritten by risk-sensitive foraging theory. In Stephens's model, energetic gains have diminishing marginal returns in fitness due to the workings of the threshold value that energy reserves have to exceed by nightfall in order to avoid starvation overnight. That is, a given amount of energy will matter more to a bird that is close to starvation than a well-fed specimen. Aktipis and Kurzban suggest that this biological mechanism supports the risk aversion for gains postulated by prospect theory. McDermott et al. (2008) go one step further and explicitly identify the energy threshold value in the daily energy budget rule with the reference point in prospect theory. They assert that risk seeking is optimal from an evolutionary perspective in the domain of losses, where a forager expects an energetic shortfall compared to the energy threshold value that guarantees overnight survival. Further, they assert that risk aversion is optimal in the domain of gains. That is, being risk averse maximizes the probability of surviving to the next day when the forager expects to exceed the energy threshold in Stephens's model.

Houston et al. (2014) critically analyze the relationship between risk-sensitive foraging theory and prospect theory. They highlight that the formal connection between risk-sensitive foraging theory and prospect theory established by McDermott et al. (2008) is only valid under rather restrictive assumptions, such as the forager having no choice between foraging options, there is no benefit of building up excess reserves above the critical energy threshold for overnight survival and there are no upper or lower boundaries on energy reserves. Furthermore, Houston et al. argue that the threshold value in the daily energy budget rule cannot be identified with the reference point in prospect theory as suggested by McDermott et al.

Setting these criticisms aside, I will develop a further critique of the use of risk-sensitive foraging models for explaining prospect theory risk preferences. A recent evolutionary model of prospect theory preferences drawing on insights from behavioural ecology is provided by Mallpress et al. (2015). In line with risk-sensitive foraging theory, the model assumes that nature selects for strategies that maximize the reproductive value of a forager. In the model, reproductive value crucially depends on the energy reserves of an agent. In particular, it is assumed that a forager can only reproduce if the organism builds up sufficient energy reserves. If the forager's energy reserves reach (or overshoot) a given threshold, the forager reproduces and gains a fixed fitness payoff in terms of reproductive units but also loses a particular amount of energetic reserves. The forager then continues at this new energy reserve level and can reproduce again if it acquires a sufficient amount of energy reserves until it dies (i.e. its energy reserve level reaches zero). The energy reserves of the forager are affected by the state of the environment. In some environmental states the energy reserves increase while in others the reserves decrease. It is assumed that environmental states change stochastically over time and the pattern of change shows auto-correlation.

Given these assumptions about the environment and the reproductive mechanism of a forager, Mallpress et al. investigate the fitness impact of a hypothetical gamble that involves choosing between the deterministic background rate of energetic gain in a given environment and a stochastic option of energy acquisition. They demonstrate that the fourfold pattern of risk preferences over changes in energy reserves enhances fitness in a variety of stochastic environments showing intermediate degrees of auto-correlation. However, the fourfold pattern ceases to be optimal and universal risk aversion is selected for when the mean change in energy reserves across the possible environmental states is positive.

Mallpress et al. demonstrate that under certain environmental conditions the fourfold pattern of prospect theory with regard to energy reserves is selected for. How does this explain prospect theory preferences over monetary lotteries shown in some experimental studies? An explanatory strategy invoked by Okasha (2007) is to postulate a currency shift from offspring numbers to money in his adaptationist explanation of risk aversion. The justification of such a move is typically that offspring numbers in biological models share a number of money-like features. In a similar vein, a currency shift from energy to money can be postulated. Of course, the deterministic link between energy reserves and reproduction assumed in Mallpress et al. does not hold when energy reserves are substituted by monetary wealth and the model is applied to contemporary western societies. Humans typically do not reproduce once their bank account surpasses a certain threshold value.[7] But suppose one accepts that there is a close link between money and energy. What are the implications of this explanatory strategy?

According to Smith et al., human behavioural ecologists identify conditional strategies of the form "in environment X, do α" and "in environment Y, do β" (2001, 112). Mallpress et al. show that prospect theory preferences over energy gambles (denoted as action α) result from an environment (denoted as environment X) in which the mean change in energy reserves across environmental states is (approximately) zero. In contrast, risk averse behaviour over energy gambles (denoted as β) in both the gains and loss domain is selected for in a situation in which the mean change in energy reserves is positive (denoted as Y). By applying the currency shift from energy to money, situation X translates into an environment X* in which the mean change in monetary wealth across states of the world is zero while situation Y translates into an environment Y* in which the mean change in wealth is positive. Similarly, prospect theory preferences over energy gambles α translate into prospect theory preferences over monetary gambles α* while risk aversion with regard to energy gambles β translates into risk aversion with regard to monetary lotteries β*. In situation X* prospect theory preferences over money α* are fitness enhancing, while in situation Y* risk averse preferences over money β* are selected for.

In combination with the currency shift from energy to money, the evolutionary model of Mallpress et al. then establishes explanations of the form "If situation X* holds, then risk preferences α* are optimal". In order to assess whether this conditional can account for human risk-taking behaviour observed in experimental monetary gambles, the assumptions embodied in the antecedent condition X* have to be checked. That is, one has to assess the degree of auto-correlation between choices and the extent to which current options allow to make inferences regarding the availability of future options.

Mallpress et al. are frank in admitting that the conditions of their evolutionary model are typically not met by the experimental set-ups in

[7] Grüne-Yanoff (2011) raises a similar point in his discussion of the use of evolutionary game theory in the social sciences. He argues that while animals largely exist on the subsistence level, humans mostly do not. As a consequence, it is much less clear what the implications of the compliance with conventions or norms are for survival and reproduction in humans compared to the implications for survival and reproduction in non-human animals.

studies of human decision making. So, how does the model explain human risk-taking behaviour? Mallpress et al. suggest that in studies on human decision making, "people may be acting on the basis of evolved predispositions that are adapted to natural environments with a richer temporal structure" (2015, 369).

If our attitudes towards risk are adapted to an environment that has a richer temporal structure than the present one (e.g., by environmental change showing a certain degree of auto-correlation), then the view of Mallpress et al. stands in conflict with the methodological assumption of human behavioural ecology that humans act optimally in their present environment. Their position here shares similarities with mainstream evolutionary psychology, which postulates that complex human behavioural traits are adapted to an ancestral environment that differs significantly from the present one. Following this line of reasoning, Mallpress et al. seem to have two options. The first option adapts the view of evolutionary psychology that there is an ancestral environment, typically seen as the Pleistocene environment inhabited by our hunter-gatherer ancestors that shaped human attitudes towards risk. Mallpress et al. would have then to make the case that this environment had a particular stochastic structure, say, show a certain degree of auto-correlation, in order to make the case for the evolution of the fourfold pattern of risk preferences. Mallpress et al. gesture at this option by pointing out that most environments, including those in which our human ancestors evolved, show some degree of auto-correlation. This option, however, runs into the difficulty that the evidence speaks against the universality of the fourfold pattern of risk preferences as discussed in section 3.

The second option allows for a variety of different ancestral environmental conditions some of which favoured the evolution of prospect theory risk preferences while others selected for risk aversion. While this option allows for a plethora of evolved human attitudes towards risk, it does not offer a rationale for the observed diversity in human decision making involving monetary gambles. For instance, one might ask: Under which condition should we expect to see experimental subjects show the fourfold pattern of risk preferences? And, under which conditions do experimental subjects show risk aversion? A natural answer to these questions would be to refer to the conditions described by situations X* and Y*, respectively. However, Mallpress et al. make it clear that this is not their explanatory strategy when they point out that conditions such as X* and Y* are typically not met in experimental tests of human decision making under risk. This leaves the problem of identifying the conditions under which different evolved risk-taking behaviour is to be observed in monetary gambles unaddressed. Phrased differently, it is left unclear what triggers an evolved predisposition towards risk-taking. Without this further detail, however, it is difficult to see whether Mallpress et al. are on the right track with their proposed model. I therefore suggest that the evolutionary model of Mallpress et al. offers only an incomplete account of human attitudes towards risk. A further explanatory step is needed that bridges the gap between the evolution of risk attitudes in ancestral environments and the risk-taking behaviour in experimental studies of decision making involving monetary lotteries.

VI. RISK AND CULTURAL EVOLUTION

While my previous remarks have been mainly critical in character, this does not imply that I reject evolutionary thinking about risk preferences tout court. In this section, I would like to widen the scope and discuss some evolutionary approaches to human behaviour by drawing on ideas from cultural evolution. Doing so goes along with a shift of gear. Rather than assessing particular evolutionary models of risk preferences in detail, I will offer some suggestive remarks on what the literature on cultural evolution can contribute to our understanding of human attitudes towards risk.

Cultural evolution refers to the change in socially transmitted beliefs, customs, skills, preferences and languages. A number of theories of cultural evolution have been proposed in biology and the social sciences. Richerson and Boyd (2005), for instance, develop formal evolutionary models to explain how human populations have changed over time under the influence of various forms of learning. By augmenting standard evolutionary models of population change with social learning processes such as imitation and teaching, they exploit the fact that learning allows human populations to change very quickly and to adapt to their environment without the workings of natural selection. The question of whether these learning processes are similar to those at play in biological evolution is only of secondary importance in Richerson and Boyd's work. As such, their work differs from what Lewens (2015) calls the ‘selectionist approach’ to cultural evolution, which maintains that cultural items such as ideas, tools and practices compete in a Darwinian struggle for survival. Proponents of the selectionist approach, such as Mesoudi (2011), suggest that cultural change can be described as a Darwinian evolutionary process that is similar in key respects to biological evolution.

A variety of non-genetic transmission processes can shape human preferences. Religious attitudes and political preferences, for instance, are typically learned from the parents while clothing preferences are strongly influenced by one's peers. Furthermore, non-peers and non-parents, such as teachers and grandparents, can shape our attitudes and preferences (Cavalli-Sforza and Feldman 1981). Independent of whether these transmission mechanisms can be understood in Darwinian terms, there exist good reasons to reflect on the role of these learning processes when accounting for the evolution of risk preferences. Dohmen et al. (2012), for instance, provide evidence for both the transmission of risk attitudes from parents to children and the influence of other role models in the environment on child risk attitudes. In addition, Dohmen et al. make the case that the transmission of risk attitudes from parents to children cannot be reduced to solely genetic factors but require also some form of socialization. For example, they observe that children reproduce the specific variation in attitudes across contexts observed in the parents and argue that this phenomenon is hard to explain with genetics and indicates that socialization is a rather fine-tuned process. As a consequence, ignoring non-genetic transmission processes may result in leaving out some potentially important preference forming mechanisms.

Cultural evolution theorists, however, have not studied the evolution of risk preferences in detail. A notable exception is Stern (2010), who studies the evolution of risk preferences by means of a biological model that includes both a genetic inheritance mechanism and a non-genetic form of inheritance of a parent's experience. He interprets this non-genetic transmission mechanism by reference to the inheritance of property and acquired knowledge commonly found in the human population. Taking into account forms of ‘cultural inheritance’, such as property and acquired knowledge, can only be seen as a first step towards a more comprehensive treatment of the coevolution of genes and culture that lead to the presently observed human attitudes towards risk.

VI. CONCLUSION

While any final verdict on evolutionary explanations of risk preferences would be premature, some general comments on the prospects and challenges of such explanations are in order. The previous discussion has focused on the fourfold pattern of risk preferences postulated by prospect theory as the explanandum of an evolutionary explanation. While this step was motivated by the prominent status of prospect theory as a descriptive account of decision making under risk, doing so led to a rather sceptical conclusion with regard to the possibility of explaining these preferences by means of evolutionary psychology understood as an evolutionary account of universal human traits. Matters would be different, however, if a feature of human decision making is selected as the target of an evolutionary explanation that has better empirical support than the fourfold pattern of risk preferences.

Returning to Harrison and Swarthout's study, a concave utility function is estimated for both expected utility theory as well as rank-dependent utility theory that emerges as the best performing non-expected utility theory. This suggests that a concave utility function, representing diminishing marginal returns of wealth, constitutes a more promising candidate for a universal feature of human preferences. As such, a concave utility function is a more suitable phenomenon to be explained by mainstream evolutionary psychology. Assuming a currency shift between monetary wealth and food, there is a plausible biological rationale for a concave utility function since reproductive output frequently scales concavely with food intake, that is, additional food leads to additional offspring but it does so with diminishing returns. Indeed, fitness functions of this kind are regularly invoked in risk-sensitive foraging theory (Houston and McNamara 1999).

Of course, having established a concave utility function does not by itself specify how agents make decisions under risk. For instance, it remains to be answered whether or not agents assign particular weights to the probabilities in their decision making process as suggested by rank-dependent utility theory. Phrased differently, the additional question arises of whether agents apply expected utility theory or some form of non-expected utility theory. Another lesson to draw from Harrison and Swarthout's study is that human decision making under risk is heterogeneous. While most of their experimental subjects apply rank-dependent utility theory, a smaller group makes decisions in line with expected utility theory. An adequate explanation of human risk attitudes has to provide a rationale for the apparent diversity in probability weighting. It cannot be presumed that a single decision theoretic procedure has become fixed in the human population.

While human behavioural ecology rightly stresses the diversity of human behaviour, it typically focuses on the ecological conditions giving rise to diverse behavioural patterns. As a consequence, similar behavioural patterns should be observed in similar environments. With regard to human decision making under risk, however, this is not necessarily the case. In particular, it is unclear whether experimental subjects showing diverse decision making under risk can be said to operate under different local ecological conditions. Theories of cultural evolution offer a further perspective on how evolutionary thinking can contribute to our understanding of risk preferences. It remains to be seen whether taking into account non-genetic transmission processes discussed by cultural evolutionist can offer an adequate explanation of the diversity in human decision making under risk.

REFERENCES

Aktipis, Athena C., and Robert O. Kurzban. 2004. "Is Homo Economicus Extinct? Vernon Smith, Daniel Kahneman and the Evolutionary Perspective." *Advances in Austrian Economics* 7, 135-153.

Brennan, Thomas. J., and Andrew W. Lo. 2011. "The Origin of Behavior." *Quarterly Journal of Finance* 1 (1), 55-108.

Buller, David J. 2005. *Adapted Minds: Evolutionary Psychology and the Persistent Quest for Human Nature*. Cambridge, MA: MIT Press.

Buss, David M. 1992. "Mate Preferences Mechanisms: Consequences for Partner Choice and Intrasexual Competition." In *The Adapted Mind: Evolutionary Psychology and the Generation of* Culture, edited by J. H. Barkow, L. Cosmides, and J. Tooby, 249-266. New York: Oxford University Press.

Buss, David M. and Patricia H. Hawley (eds.). 2010. *The Evolution of Personality and Individual Differences: Past, Present, and Future.* New York: Oxford University Press.

Caraco, Thomas. 1980. "On Foraging Time Allocation in a Stochastic Environment." *Ecology* 61 (1), 119-128.

Cavalli-Sforza, Luigi, and Marcus Feldman. 1981. *Cultural Transmission and Evolution: A Quantitative Approach*. Princeton, NJ: Princeton University Press.

Cosmides, Lela, and James Tooby. 1997. *Evolutionary psychology: A primer.* Available from <http://www.cep.ucsb.edu/primer.html>.

Delton, Andrew W., Theresa E. Robertson, and Douglas T. Kenrick. 2006. "The Mating Game Isn't Over: A Reply to Buller's Critique of the Evolutionary Psychology of Mating." *Evolutionary Psychology* 4 (1), 262-273.

Dohmen, Thomas J., Armin Falk, David Huffman, and Uwe Sunde. 2012. "The Intergenerational Transmission of Risk and Trust Attitudes". *The Review of Economic Studies* 79 (2), 645-677.

Ellis, Bruce. J. 1992. "The Evolution of Sexual Attraction: Evaluative Mechanisms in Women." In *The Adapted Mind: Evolutionary Psychology and the Generation of Culture,* edited by J. H. Barkow, L. Cosmides, and J. Tooby, 267-288. New York: Oxford University Press.

Friedman, Daniel R., Mark Isaac, Duncan James, and Shyam Sunder. 2014. *Risky curves: On the Empirical Failure of Expected Utility*. New York: Routledge.

Gould, Stephen J., and Richard C. Lewontin. 1979. "The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme." *Proceedings of the Royal Society of London B* 205 (1161), 581-598.

Grüne-Yanoff, Till. 2011. "Evolutionary Game Theory, Interpersonal Comparisons and Natural Selection: a Dilemma." *Biology and Philosophy* 26 (5), 637-654.

Harrison, Glenn. W., and Don Ross. 2017. "The Empirical Adequacy of Cumulative Prospect Theory and its Implications for Normative Assessment." *Journal of Economic Methodology* 24 (2), 150-165.

Harrison, Glenn W., and Todd J. Swarthout. 2016. "Cumulative Prospect Theory in the Laboratory: A Reconsideration." ExCEN Georgia State University Working Paper. Available from < https://cear.gsu.edu/wp-2016_05-cumulative-prospect-theory-laboratory-reconsideration/>.

Houston, Alasdair I., Tim W. Fawcett, Dave E. W. Mallpress, and John M. McNamara. 2014. "Clarifying the Relationship Between Prospect Theory and Risk-Sensitive Foraging Theory." *Evolution and Human Behaviour* 35 (6), 502-507.

Houston, Alasdair. I., and John M. McNamara. 1999. *Models of Adaptive Behaviour.* Cambridge: Cambridge University Press.

Kitcher, Philip. 1985. *Vaulting Ambition: Sociobiology and the Quest for Human Nature*. Cambridge, MA: MIT Press.

Laland, Kevin N., and Gillian R. Brown. 2002. *Sense and Nonsense: Evolutionary Perspectives on Human Behaviour.* Oxford: Oxford University Press.

Lewens, Tim. 2015. *Cultural Evolution.* Oxford: Oxford University Press.

Machery, E. 2008. "A Plea for Human Nature." *Philosophical Psychology* 21 (3), 321-329.

Mallpress, Dave E.W., Tim W. Fawcett, Alasdair I. Houston, and John M. McNamara. 2015. "Risk Attitudes in a Changing Environment: An Evolutionary Model of the Fourfold Pattern of Risk Preferences." *Psychological Review* 122 (2), 364-375.

McDermott, Rose, James H. Fowler, and Oleg Smirnov. 2008. "On the Evolutionary Origin of Prospect Theory Preferences." *Journal of Politics* 70 (2), 335-350.

McNamara, John M., and Alasdair I. Houston. 1992. "Risk-sensitive Foraging: A Review of the Theory." *Bulletin of Mathematical Biology* 54 (2-3), 355-378.

Mesoudi, Alex. 2011. *Cultural Evolution: How Darwinian Theory Can Explain Human Culture and Synthesize the Social Sciences.* Chicago: University of Chicago Press.

Mishra, Sandeep and Laurence Fiddick. 2012. "Beyond Gains and Losses: The Effect of Need on Risky Choice in Framed Decisions." *Journal of Personality and Social Psychology* 102 (6), 1136-1147.

Okasha, Samir. 2007. "Rational Choice, Risk Aversion, and Evolution." *Journal of Philosophy* 104 (5), 217-235.

Quiggin, John. 1982. "A Theory of Anticipated Utility." *Journal of Economic Behavior and Organization* 3 (4), 323-343.

Richerson, Peter J., and Robert Boyd. 2005. *Not by Genes Alone: How Culture Transformed Human Evolution.* Chicago: University of Chicago Press.

Schulz, Armin W. 2008. "Risky Business: Evolutionary Theory and Human Attitudes toward Risk—A Reply to Okasha." *Journal of Philosophy* 105 (3), 156-165.

Smith, Eric A., Monique Borgerhoff Mulder, and Kim Hill. 2001. "Controversies in the Evolutionary Social Sciences: a Guide for the Perplexed." *Trends in Ecology and Evolution* 16 (3), 128-135.

Stephens, David W. 1981. "The Logic of Risk-sensitive Foraging Preferences." *Animal Behaviour* 29, 628-629.

Stern, Michael D. 2010. "Patrimony and the Evolution of Risk Taking." *PLoSONE* 5 (7), e11656.

Suppes, Patrick, David H. Krantz, Duncan Luce, and Amos Tversky. 1989. *Foundations of Measurement, Vol. 2.* San Diego: Academic Press.

Tooby, James, and Lela Cosmides. 1990. "On the Universality of Human Nature and the Uniqueness of the Individual: The Role of Genetics and Adaptation." *Journal of Personality* 58 (1), 17-67.

Tversky, Amos, and Daniel Kahneman. 1992. "Advances in Prospect Theory: Cumulative Representation of Uncertainty." *Journal of Risk and Uncertainty* 5 (4), 297-323.

Wilcox, Nathaniel T. 2008. "Stochastic Choice Models for Binary Discrete Choice Under Risk: A Critical Primer and Econometric Comparison." In *Risk Aversion in Experiments. Research in Experimental Economics, Vol. 12*, edited by J. Cox and G. W. Harrison, 197-292. Bingley: Emerald.

Bengt Autzen is a postdoctoral fellow at the Munich Center for Mathematical Philosophy having received a PhD in philosophy from the London School of Economics (LSE). His research interests lie in the philosophy of biology, the philosophy of statistics and the philosophy of social sciences.
Contact e-mail: <Bengt.Autzen@lrz.uni-muenchen.de>

Erasmus Journal for Philosophy and Economics,
Volume 11, Issue 1,
Spring 2018, pp. 50-79.
https://doi.org/10.23941/ejpe.v11i1.246

Naturalism and Moral Conventionalism: A Critical Appraisal of Binmore's Account of Fairness

CYRIL HÉDOIN

REGARDS Research Center, University of Reims Champagne-Ardenne

Abstract: This article provides a critical examination of Ken Binmore's theory of the social contract in light of philosophical discussions about moral naturalism and moral conventionalism. Binmore's account builds on the popular philosophical device of the *original position* but gives it a naturalistic twist. I argue that this makes it vulnerable to moral skepticism. I explore a possible answer to the moral skeptic's challenge, building on the fact that Binmore's account displays a variant of moral conventionalism. I ultimately conclude however that the conventionalist answer leads to a purely behaviorist view of morality, which implies that there is nothing special about morality and fairness norms. I propose alternative interpretations of conventionalism. These accounts escape most of the difficulties because they place emphasis on the reasons that establish a moral convention.

Keywords: Binmore, moral naturalism, moral conventionalism, original position, fairness

JEL Classification: B00, B31, C73, D02

1. INTRODUCTION

Economists have demonstrated during the last three decades a growing interest in issues related to fairness and morality. Indeed, the rise of game theory has considerably changed the disciplinary landscape between economics and moral philosophy: economists now have a tool at their disposal directly relevant to making significant contributions to moral philosophy. This article provides a critical examination of a specific attempt to produce a theory of fairness through the game-theoretic lens, namely Ken Binmore's theory of the social contract

AUTHOR'S NOTE: I would like to thank the two anonymous referees for their challenging and very useful comments on the previous version of this article. I also thank O. Çağlar Dede for his editorial work. All errors and omissions are of course mine.

(Binmore 1994, 1998, 2005). Binmore presents his account as an attempt to “treat morality as a science” (Binmore 2005, 1). It pursues two goals: first, to account for the origins and the content of our fairness judgments; second, to argue for an egalitarian view of fairness. Clearly, the justifiability of the second prescriptive goal depends on the success of the first descriptive goal. However, several philosophers have argued that pursuing the first goal might undermine the justifiability of the second (see, for instance, Joyce 2006). My examination of Binmore’s account responds to this general philosophical worry.

Binmore’s theory of fairness builds on the popular philosophical device of the *original position*, independently developed by John Rawls (1971) and John Harsanyi (1953). However, Binmore gives a naturalistic twist to this device. He naturalizes it through two related claims: first, he argues that genetic and biological evolution has encoded the original position in our genes. In particular, he claims, biological evolution has endowed us with the ability to sympathize and empathize with others, regardless of genetic relatedness. Second, Binmore argues that cultural evolution has led to the emergence of standards of fairness under the forms of empathetic preferences that make interpersonal comparisons of utility possible. The original position is then conceived by Binmore as a genetically encoded but culturally loaded algorithm, which humans use to coordinate in the “game of life”, i.e. the game whose “rules are determined by the laws of physics and biology; by geographical and demographic facts; by technological and physiological constraints” (Binmore 1998, 4). The game of life has a multiplicity of Pareto-efficient equilibria. The original position device is instantiated in what Binmore calls the ‘game of morals’ and selects one equilibrium on the basis of an egalitarian standard of fairness.

My goal in this paper is to clarify the implications of the naturalization of the original position for the status and the significance of fairness claims and judgments. I shall argue that the means by which the original position is naturalized makes it vulnerable to moral skepticism. Specifically, I argue that Binmore’s naturalization of the original position implies that fairness judgments are grounded on the power structure of the society. A moral skeptic can then argue that these judgments do not have any moral content and authority, and thus, cannot be objectively true. I explore a possible answer to the moral skeptic’s challenge by arguing that Binmore’s account displays a variant of moral conventionalism. However, I conclude that Binmore’s conventionalist answer leads to a purely behaviorist view of morality, which implies that there is nothing special about morality and fairness norms. In response, I consider alternative accounts of moral conventionalism which emphasize the importance of the reasons that establish moral conventions. These alternatives escape most of the difficulties which are associated with Binmore's account.

The article is organized as follows. Section 2 presents Binmore's account by explaining the naturalization of the original position as a device to coordinate in the game of life. Section 3 raises the critique from moral skepticism against Binmore's account, as the latter is understood as an instance of moral naturalism. Section 4 examines a possible answer to this critique by characterizing Binmore's account as an instance of moral conventionalism. Section 5 argues that Binmore's moral conventionalism nevertheless fails to answer the skeptic's critique, while also demonstrating that other forms of moral conventionalism are immune to it.

2. BINMORE'S NATURALISTIC ACCOUNT OF FAIRNESS

Ken Binmore's naturalistic account of the social contract and fairness norms is developed in a two-volume book *Game Theory and the Social Contract* (Binmore 1994, 1998).[1] It builds on the ideas of three influential authors in moral and political philosophy: David Hume, John Harsanyi and John Rawls. It ultimately leads to a vindication of Rawls's egalitarianism against Harsanyi's utilitarianism. Binmore sees in Hume the foundations of a 'conventionalist' view of justice in which fairness norms are taken to be the product of an evolutionary process. While they are initially conceived as conventional devices to solve coordination problems, fairness norms progressively acquire normative power as they become the commonly understood standard for determining whether a situation is just or unjust. From Rawls and Harsanyi, Binmore's account retains the philosophical concept of the *original position* that both authors simultaneously developed in the 1950s.

Binmore's naturalistic account of the social contract is in line with the growing body of scholarship that applies tools and models from

[1] Binmore exposes his account in a less mathematical and less detailed fashion in a later book, *Natural Justice* (2005). This book does not add anything substantive to the preceding two volumes except for its more straightforward presentation of the main ideas. Therefore, I will not refer to it except in the few instances where it indicates that Binmore has changed his mind with regard to what is written in *Game Theory and the Social Contract* (1998).

natural and social sciences to issues of moral and political philosophy. More specifically, it is a representative contribution of a set of approaches combining the mathematical principles and concepts of game theory with theories of natural and cultural evolution to study the origins of morality.[2] In the present case, Binmore's naturalism develops as an attempt to naturalize Rawls's and Harsanyi's original position (henceforth, OP). The term 'naturalize' and the derivative 'naturalization' here refer to the fact that Binmore attempts to show that the OP is not merely a philosophical thought experiment. It is actually part of the natural world in the sense that it corresponds to a device—or an algorithm—that humans are using to solve coordination problems. Indeed, Binmore argues that the device of the OP is actually a genetically-encoded algorithm used by people to make fairness judgments because of our natural history. Moreover, the use of the OP algorithm depends on standards for making interpersonal comparisons of utility that evolve from *cultural* evolutionary processes. Binmore substantiates these claims through a game-theoretic formalization of the bargaining that takes place behind the veil of ignorance constitutive of the OP.

Binmore defines a social contract as "the set of common understandings that allow the citizens of a society to coordinate their efforts" (2005, 3). He claims that any social contract must satisfy three requirements: stability, efficiency and fairness. The first is the most important. Binmore rejects ad hoc assumptions that moral philosophers have sometimes invoked to make the agreement concluded under the veil of ignorance binding (see, for example, Gauthier 1986). Since every member of a society is part of the social contract (including government members and law enforcers), a stable agreement must be self-enforcing. Arguing that any social contract relies on a *repeated* game, Binmore makes use of the folk theorem of repeated game theory according to which multiple equilibria exist across all different sorts of games as soon as a given game is indefinitely repeated between the same players. The folk theorem shows that the evolution of cooperation is not dependent on the existence of prosocial preferences. Several stable social contracts are then possible, without necessarily depending on prosocial preferences.

[2] Other representative works include Alexander (2010), Skyrms (1996, 2004), Sugden (2005), and Young (2001).

Next to the stability requirement, the two other conditions for the viability of a social contract are efficiency and fairness. Efficiency is defined by the Pareto criterion and is based on a simple argument about group selection. Different communities may agree on different social contracts. If we assume that communities' expansion is a function of the efficiency of their social contracts, then communities operating under a sub-optimal one will progressively be trumped over by those operating with efficient ones.[3] If we assume, like Binmore, that the negotiation of a social contract takes the form of a bargaining game between two players, then any viable social contract is contained in the area between the disagreement point (which corresponds to the minimax gain for each player) and the maximum that each player can gain. An efficient agreement is by definition placed on the Pareto frontier, delimiting the set of feasible social contracts. However, numerous agreements are still possible. According to Binmore, the fairness criterion serves as a device to select one of the efficient equilibria. Fairness norms, therefore, help individuals to coordinate on a particular outcome through the expectation that everyone will choose it. Since individuals agree to coordinate on a particular equilibrium, it is deemed to be fair in a sense that I will now explain.

How the fair equilibrium is determined, and thus which equilibrium will be chosen, are the central questions answered by Binmore's naturalistic account of justice. This lays the groundwork for the naturalization of the OP. According to Binmore, humans are engaged in an ancestral 'game of life' whose rules are defined by biological constraints. Binmore describes it as a (repeated) bargaining game played by two players, Adam and Eve. He takes the problem of food-sharing in human foraging communities as the hallmark problem encoded in the game of life. From the very start of human history, sharing food has been an allocation problem that any community had to solve. Since the same problem is also faced by animals such as chimpanzees, Binmore makes the conjecture that humans have been genetically programmed to

[3] As Binmore notes, this argument is immunized against the traditional critiques of group selection explanations (1998, 185). In fact, the stability requirement assures that any existing social contract is an equilibrium and therefore that individuals have an interest to enforce it. Sugden (2001b) notes however that Binmore does not link reproduction with utility. As in evolutionary game-theoretic models in economics, utility describes only the propensity for a strategy to be replicated in the society. If Pareto optimality is defined in terms of utility, then assuming that Pareto improvements promote survival or competitiveness is a *non sequitur*. Some behavioral patterns with strong replication propensity, such as addictive behaviors, can be destructive in the long run.

play the game of life. In the context of genetically related individuals, it is easy to show that sharing food with one's relatives is an equilibrium (not the only one, however), which might be selected and implemented through a food-sharing insurance contract. Under such a contract, unlucky relatives who have failed to get any food receive some food from more lucky relatives. Indeed, this kind of mechanism makes it more likely that genes shared by relatives will spread. However, in human societies, cooperation expands well beyond the circle of genetic relatives: food-sharing insurance contracts also take place in the context of genetically unrelated individuals facing uncertainty about the results of their hunt. Under this 'veil of uncertainty' where one does not know whether he will be able to catch any food in the future, each person must sympathize with her possible future selves ('Mr. Lucky' or 'Ms. Unlucky') by anticipating how much their future preferences would be satisfied in the different possible scenarios. A food-sharing insurance contract represents a Pareto-improvement for agents facing such kind of uncertainty. Moreover, the folk theorem of repeated game theory indicates that such contracts are sustainable as equilibria. Binmore contends that the device of the OP first evolved as a way to negotiate such contracts (1998, 219). On this basis, it progressively became a genetically-encoded algorithm used to solve more general and larger fairness issues:

> people take a technique used within one circle of social problems and unthinkingly apply it to a wider domain of problems. In so doing, they continue to play by the rules of the game for which the technique originally evolved, not noticing—or pretending not to notice—that the rules of the game played in the wider circle may be quite different (Binmore 1998, 219).

There is, however, a clear difference between the veil of uncertainty that members of hunter-gatherer groups had to deal with and the veil of ignorance of the OP as it was initially conceived by Harsanyi (1953): in the latter, persons under the veil of ignorance have to put themselves in *others*' position to determine what they have to do. They must *empathize* with other persons by pretending that they have *the other's* preferences; they must assume that they literally *are* these other persons. This implies that each person has the ability to make comparisons (either at the level of utility or the level of preference satisfaction) between each other member of the population. According to Binmore, the use of the OP as a device to make fairness judgments thus evolved from the combination of food-sharing negotiations between genetically unrelated individuals and interactions between genetically related individuals:

> all that is then needed is for us to hybridize these two processes by allowing a player to replace one of the future persons that a roll of dice might reveal him to be, by a person in another body who is to be treated in much the same way that he treats his sisters, his cousins or his aunts (Binmore 1998, 220).

Cooperation with non-relatives through the OP is thus partially the product of natural evolution: first, kin selection has 'programmed' organisms to cooperate with genetic relatives; second, natural pressures due to the uncertainty regarding feeding in hunter-gatherer societies have favored the selection of an innate ability to empathize with others. However, while natural selection has endowed us with an innate ability to make fairness judgments through the OP device, it has not determined the *content* of these judgments. Following Harsanyi, Binmore considers that individuals possess *empathetic preferences* allowing them to determine if they prefer to be 'person *i* in situation *x*' or 'person *j* in situation *y*'. These empathetic preferences make the agents' utility functions commensurable and determine the rate at which the utilities of each individual will be traded with those of others (Binmore speaks of 'social indices'). The content of empathetic preferences and therefore the value of the social indices are determined by cultural evolution.[4] On this basis, the fairness of the social contract is established by adding a device to the game of life through which individuals will be able to coordinate their action: the 'game of morals'.

The game of morals is purely conventional. Binmore interprets it as a heuristic through which individuals reflect on and anticipate the reaction of every member of a society when a new social contract is established. Like the game of life, the game of morals leads individuals to make use of empathy. Binmore contends that any individual can at every moment appeal to the game of morals if he is not satisfied with his situation. Appealing to the game of morals is like rolling the dice again and negotiating a new social contract under a veil of ignorance.

[4] Binmore makes an analogy with the evolution of language. The capacity to develop and to learn a language has a biological origin. However, the content of any language comes from the cultural history of each society and is independent of any biological factor.

The clause of unlimited appeal has radical implications: first, it means that individuals must have the same empathetic preferences—Binmore calls it *symmetric empathetic equilibrium*. On the contrary case, an agreement would not be reached, leading agents to play the game of morals again and again. Second, the agreement must be considered fair by each individual according to the existing social indices. As Binmore puts it:

> A *fair social contract* is then taken to be an equilibrium in the game of life that calls for a use of strategies which, if used in the game of morals, would never leave a player with an incentive to exercise his right of appeal to the device of the original position [...] the game of morals is nothing more than a coordination device for selecting one of the equilibria in the game of life (2005, 172, his emphasis)

Without the existence of any enforcing authority, Binmore shows that the agreed social contract will correspond to the proportional (or egalitarian) bargaining solution. The solution ensures that the players' utility functions are suitably rescaled according to the social indices that correspond to the prevailing empathetic equilibrium. Indeed, it is clear that any social contract which is not egalitarian will lead the worst-off individuals to appeal for a new game of morals. If everyone uses the game of morals to choose a fair social contract, and if this becomes common knowledge, then this necessitates an egalitarian social contract.

While Binmore is rather vague about the precise mechanisms responsible for the evolution of empathetic preferences,[5] he provides an interesting argument for their role in determining the equilibrium reached in the game of morals. The argument is somewhat complex but relies on two key assumptions: the first assumption is that *real* bargains always converge toward the Nash solution; the second assumption is that the enforcement of the agreement reached behind the veil of ignorance is ultimately always self-enforcing (for example, there is no external enforcer). The former is indeed essential in the algorithm Binmore proposes to compute the value of the "social indices" (Binmore 1998, 441). He considers three temporal scales in the evolution of fairness norms (1998, 226-227). In the *short run*, both individuals' personal and empathetic preferences are fixed and only their choices made through the device of the OP can change. In the *medium run*, the

[5] Binmore had initially made use of Richard Dawkins's (1975) concept of 'meme' to account for cultural revolution. He has retreated, however, in *Natural Justice*, acknowledging the huge difficulties related to this concept.

individuals' personal preferences are fixed, but their empathetic preferences are susceptible to change, as the Pareto-frontier of the bargaining game they are playing moves. In the *long run*, both empathetic and personal preferences can change through the forces of cultural evolution and biological evolution respectively. The temporal scale of cultural evolution is thus the medium run, while biological evolution operates in the long run. In the short run, the agents play the game of morals—using the OP device to select one of the Pareto-efficient equilibria. The result of the bargain is perceived as 'fair' by the participants because it selects the egalitarian equilibrium *given the (symmetric) empathetic preferences* of the players. However, these empathetic preferences find their origins in the bargaining relationships that take place in the game of life. In the latter, the players only rely on their bargaining skills and it is assumed that the resulting outcome corresponds to the Nash bargaining solution. As is well-known, to which point on the contract curve this solution corresponds to depends on the players' risk and/or time preferences. Relatively risk-tolerant and patient players will have an advantage in the bargaining process and will obtain the lion's share of the available resources. What happens then is that empathetic preferences are set such that the outcome corresponding to the Nash solution is selected as the egalitarian solution in the game of morals. In other words, the equilibrium in the game of morals corresponds to the egalitarian solution, taking the players' empathetic preferences to measure utility. However, at the equilibrium in the game of life, the players' utilities determined on the basis of their personal preferences are such that the Nash solution is satisfied. Obviously, the latter need not be egalitarian.[6]

Actually, Binmore's claim seems to be that empathetic preferences serve as *a posteriori* egalitarian rationalization of previous bargaining arrangements reached in the game of life. These arrangements need not be egalitarian (this depends on the players' time preferences or risk preferences, as well as their bargaining abilities), but in the context of the game of morals they must be seen as *fair* by the players; otherwise, one of them would want to 'roll the dice' again. This explains Binmore's conclusion that existing social contracts must be egalitarian when evaluated according to the players' empathetic preferences.

[6] Moreover, the very notion of 'equality' in the context of the Nash solution is meaningless since the latter does not assume that interpersonal comparisons of utility are possible.

3. FROM DESCRIPTIVE TO PRESCRIPTIVE ETHICS: NATURALISM AND MORAL SKEPTICISM

The preceding section has shown that Binmore's naturalistic account of fairness leads to a substantive moral conclusion. Assuming that the OP is a device that has been historically used to solve coordination problems in the game of life, social contracts must actually be egalitarian, at least when judged according to the prevailing empathetic preferences in the population. In this section, however, I argue that independently of what one may think of its substantive conclusion, this account has to face the same meta-ethical challenge that confronts all forms of moral naturalism. This challenge corresponds to what is generally labeled 'moral skepticism': the view according to which the naturalistic foundations of morality raise doubts about the justification of moral judgments.

The naturalization of the OP is used by Binmore as part of the larger project of treating "morality as a science" (2005, 1). However, the naturalistic project of treating morality as a science may have several meanings. The most modest interpretation is restricted to the domain of what is sometimes called 'descriptive ethics'. Naturalism then corresponds to the general endeavor of providing a scientific account of moral practices and institutions and the genealogy of moral judgments. As it will be clearer below, such a form of naturalism does not imply any commitment regarding either the existence of moral facts or properties, or the truth-value of moral beliefs. Two stronger forms of naturalism intend to provide some articulation between descriptive ethics and the domain of 'prescriptive' ethics. Indeed, they suggest that knowledge about the way moral judgments and practices evolved vindicate these judgments and practices, and more generally, moral theories. In other words, they are committed to the claim that the naturalistic origins of moral judgments and practices *justify* these judgments and practices in some well-defined sense. The first of these two forms of naturalism, *moral naturalism*, builds upon the postulate that moral properties and facts can ultimately be reduced to naturalistic properties and facts. The second form, *moral conventionalism*, takes morality to consist of nothing but conventional practices. As Binmore uses his naturalistic account as a defense of some version of Rawlsian egalitarianism, it is clear that it should be regarded as an exercise in both descriptive and prescriptive ethics, either as an instance of moral naturalism or of moral conventionalism. I focus in this section on the objections made to moral naturalism as they will prove useful to discuss moral conventionalism in the next two sections.

The main objections made against moral naturalism can be summarized in the following way: by showing that fairness and, more generally, morality have naturalistic foundations, naturalistic approaches undermine the very ground on which the normative force of morality and fairness are built upon. This 'very ground' is constituted by the naturalistic origins of morality. Far from vindicating morality, these origins make it illusory or even non-existent. This objection in particular has been made against *evolutionary* moral naturalism (that is, the set of views according to which moral values and obligations are grounded by facts about biological evolution); but it is also relevant to other forms of moral naturalism (Joyce 2006). I shall argue in this section that the objection is even more compelling with respect to Binmore's naturalization of the OP. This leads to the following problem: if Binmore's account is empirically relevant, then this leads to doubt about the moral force of fairness norms. More precisely, once one knows and accepts Binmore's account of fairness norms, then it is not clear why one should maintain that his beliefs about what is fair are justified.

Joyce (2006) develops a strong argument that moral naturalism leads to moral skepticism, the meta-ethical view according to which it is doubtful that our moral judgments and beliefs can ever be justified (Sinnot-Armstrong 2015). In the specific case of evolutionary moral naturalism, Joyce's main point is that the empirical knowledge of the genealogy of our moral judgments and beliefs (the fact that these judgments and beliefs emanate from dispositions that have evolved through natural selection) fails to justify them. The reason is that this knowledge does not entail any confidence in the idea that natural selection is likely to have produced true beliefs. As a consequence, this knowledge should lead to moral skepticism, or even to moral nihilism.[7] Consider the following analogy:

> Suppose that there were a pill that makes you believe that Napoleon won Waterloo, and another that makes you believe that he lost. Suppose also that there were an antidote that can be taken for either

[7] A moral nihilist argues that the empirical knowledge of the genealogy of moral beliefs render them *unjustifiable*—rather than merely failing to provide a justification. In this case, it is contended that we cannot provide any justification for our moral beliefs ever and therefore that we should not accept any moral claim. An obvious implication is that nothing can be morally wrong according to a moral nihilist.

> pill. Now imagine that you are proceeding through life happily believing that Napoleon lost Waterloo (as, indeed, you are), and then you discover that at some point in your past someone slipped you a 'Napoleon lost Waterloo' pill… Should this undermine your faith in your belief that Napoleon lost Waterloo? (Joyce 2006, 179).

Joyce argues—quite reasonably—that the answer to this last question should be 'yes'. Correspondingly, your knowledge of the genealogy of your belief 'Napoleon lost Waterloo' should encourage you to take the antidote. Now, if we substitute the belief 'Napoleon lost Waterloo' for any moral belief or judgment and the belief pills for natural selection, then the moral skeptic's argument is easy to understand:

> Were it not for a certain social ancestry affecting our biology, the argument goes, we wouldn't have concepts like *obligation*, *virtue*, *property*, *desert*, and *fairness* at all. If the analogy is reasonable, therefore, it would appear that once we become aware of this genealogy of morals we should (epistemically) do something analogous to taking the antidote pill: cultivate agnosticism regarding all positive beliefs involving these concepts until we find some solid evidence either for or against them (Joyce 2006, 181, emphasis in original).

The analogy works on the basis of the postulate that there is absolutely no reason to think that natural selection is likely to have produced *true* beliefs. Assuming that there are independent moral facts or that moral facts can be reduced to non-moral facts, descriptive evolutionary ethics (the scientific works examining to what extent human morality is the product of natural selection) does not provide a basis for believing that our beliefs about these facts are true. Quite the contrary, moral skepticism argues that descriptive evolutionary ethics undermines morality. Knowing the non-moral genealogy of our moral beliefs can only foster doubt about their possible truth. Moreover, moral skeptics argue that it is implausible to find in the non-moral genealogy of moral beliefs any source for the *necessary practical authority* of moral prescriptions (Joyce 2006, 190-9). In a nutshell, even though natural selection may have led to the existence of moralized social and psychological pressures R for person A to do φ, this does not imply that everything else, being equal, he *ought* to do φ. Actually, what A ought to do also depends on his desires and non-moral beliefs. In other words, it seems that there is no desire-independent practical reasoning that can endow moral beliefs with the required practical authority. Moral prescriptions would then be followed not because they are 'moral', but only because individuals have the psychological urge to conform to them due to unrelated (amoral) factors. For moral skeptics, this is a significant reason to doubt that our moral beliefs are justified.

This is not the place to pursue the issue of the plausibility of moral naturalism or skeptical critiques further. However, by introducing the argument from moral skepticism against moral naturalism, I want to show that it is directly relevant for Binmore's naturalistic account of fairness. Indeed, this argument can be reconstructed as follows: if Binmore's account is empirically relevant, then we should have doubt about the moral force of fairness norms. More precisely, once one knows and accepts Binmore's account of fairness norms, then it is not clear why one should consider that his beliefs about what is fair are justified. The moral skeptic's critique seems to be even stronger in Binmore's case, because Binmore's naturalism emphasizes the essential role played by *bargaining power* in the evolution of fairness norms. To fully establish this claim, it is first required to show that Binmore seeks to provide a non-moral genealogy to our fairness claims.[8] This is not difficult since he is quite explicit about this point. The OP is a device for making fairness judgments. It has two distinct naturalistic origins: first, it evolves from a biological and genetic genealogy that starts with the family games played by our ancestors, which continues with food-sharing insurance contracts that had to be negotiated in hunter-gatherer societies. This is what Binmore calls the 'game of life'. Second, the use of the OP in concrete cases necessitates making interpersonal comparisons of utility in what Binmore calls the 'game of morals'. This depends on the existence of empathetic preferences whose content (which materializes through 'social indices') evolves through a potential cultural genealogy. The former kind of naturalistic origins obviously makes Binmore's account a target for moral skepticism. But I would argue that moral skepticism has more bite on the latter.

Indeed, as I explain in the preceding section, empathetic preferences operate as *a posteriori* rationalizations of previous bargaining results. One may then wonder whether there is anything fair in the resulting fairness norms that select among the multiple efficient equilibria in our

[8] Since Binmore is not concerned with the naturalistic foundations of morality as a whole, but only with those of our conception of fairness, the discussion will now be restricted to the latter.

daily interactions. We thus recover the skeptic's point: once you realize that moral or fairness judgments are grounded on norms that have naturalistic origins (biological and/or social), this should raise doubts about their justification. To Binmore's credit, he is not shy about this, since he explicitly acknowledges the importance his account gives to bargaining power in the evolution of fairness.[9] Bargaining power may have several origins. As indicated above, it may result from the shape of the individual's personal preferences, the latter being a function of the individual's social position (or genealogy, considering that preferences are partially genetically transmitted). It may also result from the position of the disagreement point in the bargaining game since, by construction, the Nash solution will then favor the agent with the larger initial endowment.[10]

It seems that the moral skeptic is entitled to ask whether fairness judgments resulting from such asymmetries in bargaining power should count at all as authentic justified moral beliefs.[11] Indeed, if I know that my judgment for evaluating the fairness of a situation depends on preferences that have been shaped by the power structure of society, should I give it any more credence than my belief that Napoleon lost Waterloo when I know that it results from the fact that I have taken the appropriate pill? Moreover, since fairness judgments are a kind of moral judgment, they should have the same normative and practical force as any other moral judgments. However, unless one recognizes that the power structure of the society is itself 'fair' (whatever that may mean), it is not clear why one should grant any normative significance to his fairness judgments. I may indeed honestly judge that the current situation is fair in spite of the fact that I am disadvantaged relative to others in the population (and possibly advantaged relatively to some other persons). But why should I trust my judgment and have *any*

[9] "We have to live with the unwelcome truth that the interpersonal comparisons of utility necessary to make fairness judgments meaningful are ultimately determined by the underlying balance of power" (Binmore 1998, 425).

[10] Recall that the Nash solution corresponds to the point that maximizes the product of each player's utility at the bargaining outcome minus their utility at the disagreement point.

[11] Note that the skeptic's query is left unanswered even if one assumes some external authority and thus endorses the utilitarian solution rather than the egalitarian one in Binmore's account. The point is that fairness judgments made on the basis of empathetic preferences do not express justified moral beliefs from the skeptic's point of view.

normative reason to adhere to it knowing that it results from the fact that others were indeed advantaged in the past?[12]

It appears, then, that Binmore's account of fairness is vulnerable to the moral skeptic's rebuttal of moral naturalism. The descriptive claim that fairness judgments have naturalistic origins provides a strong reason to doubt their prescriptive validity or force. At this point, a possible response is to concede and to accept the skeptic's conclusion that fairness judgments cannot be justified. However, an alternative path is available by endorsing moral conventionalism. The next two sections evaluate whether interpreting Binmore's account as an instance of moral conventionalism can escape the skeptic's conclusion.

4. A THIRD PATH: BINMORE'S ACCOUNT AS AN INSTANCE OF MORAL CONVENTIONALISM

Though the skeptic's critique of moral naturalism is powerful, it is not plausible to assume that Binmore agrees with the skeptic's conclusions, as that would make his defense of Rawlsian egalitarianism meaningless. This section, as well as the next, investigates the third path between moral naturalism and moral skepticism, which I refer to as moral conventionalism. While I find moral conventionalism a plausible and highly attractive meta-ethical stance that potentially avoids the skeptic's conclusion, I shall argue that Binmore's naturalistic account offers a variant of moral conventionalism that falls short of vindicating fairness and morality more generally. The main reason for this is the lack of reflexivity that individuals have over their empathetic preferences in Binmore's account. In this section, I provide a characterization of moral conventionalism and explain why it may answer the skeptic's critique. The next section explains why the kind of moral conventionalism endorsed by Binmore is nonetheless unsatisfactory in this regard.

Broadly speaking, moral conventionalism can be characterized as *the meta-ethical view according to which morality is conventional*. On this view, morality is constituted by conventional rules which (by definition) (i) depend on social practices, (ii) are historically contingent and (iii) are arbitrary in some sense (see, for instance, Marmor 2009). There are several variants of moral conventionalism, all of them combining in one way or another Hume's account of justice as an artificial virtue (a virtue

[12] Note that this is a pretty weak normative requirement for a moral judgment. Most moral thinkers would require for a moral judgment to give one a *decisive* reason to abide by it.

that depends on conventional rules) with David Lewis's (2002) theory of conventions (Verbeek 2008). Let me first explain the concept of conventional rules. Clarifying this concept is indeed essential to understand why moral conventionalism is unable to avoid moral skepticism. Feature (i) is intended to capture that a convention exists in some community if and only if it is actually followed. By 'being actually followed', I mean that a convention *C* necessarily corresponds to the regularity of behavior *R* that occurs in a given community *G* under a given set of circumstances *S*. Another way to characterize this property is to say that a convention is *practice-dependent*. Feature (ii) indicates that a convention *C* has emerged and evolved through a process taking place in historical time, but that an alternative convention *C'* would have emerged and evolved had historical circumstances been different at some moment. This is the same as saying that a different convention *C'* (and thus a different regularity of behavior *R'*) could have existed in the very same community *G* under the very same set of circumstances *S*. Finally, feature (iii) is that there is no categorically imperative reason for following a convention *C*. By this, I mean that there are in principle reasons for following an alternative convention *C'* in the very same set of circumstances *S*.[13] In principle, a minimal reason for following a convention *C* is that each individual expects others to follow *C*. On this basis, I propose to characterize a convention in the following way:

> A rule *C* is conventional if and only if, for a community *G* and a set of circumstances S:
> 1) *C* is practice-dependent, historically contingent and arbitrary.
> 2) There is some *k*-order of mutual belief in *G* that *C* is followed in *S*.

The first condition follows from the three features stated above. The second is needed to ensure that the behavioral regularity *R* is not the result of pure randomness but rather of *intentional* behavior. Depending on one's preferred account of conventions, the *k*-order of mutual belief may vary between first-order mutual belief and common belief (as is the

[13] Available definitions of conventions in the literature (e.g., Marmor 2009) generally do not distinguish between historical contingency and arbitrariness. I think however that it is important not to conflate these two features. Indeed, the former feature refers to the *causal origins* of conventions while the latter rather refers to the *nature* of conventions. I return to the distinction between causal and constitutive dependency of morality upon conventions below.

case in Lewis's account). It is not needed to take a position on this last issue here.

On this basis, moral conventionalism can here be precisely defined as the view according to which the conventional nature/origin of morality concerns those rules that allow persons to coordinate and to cooperate.

Moral conventionalism has been endorsed by several economists and philosophers finding their inspiration in Hume's scholarship. In addition to Binmore (1998), Skyrms (1996) and Sugden (2005) have made significant contributions by attempting to show that fairness and morality are ultimately conventional—though they largely differ in their details. Asked to answer the skeptic's critique of moral naturalism, the moral conventionalist is most likely to simply reject the skeptic's two fundamental premises that 1) moral claims must depend on justifiably true beliefs and 2) moral claims have an unconditional normative force. The conventionalist's best defense consists in denying that there are the kinds of moral facts and moral claims of the sort that the skeptic argues for: facts and claims that depend on justifiably true beliefs and have unconditional normative forces. This is not a problem for the conventionalist though: there are other kinds of facts (let's call them 'conventional facts'), and according to the conventionalist these are the sole ones that constitute morality. These facts refer to tacit and arbitrary agreements between persons that solve coordination problems. This is clearly a view espoused by Binmore who emphasizes that fairness norms have been primarily designed to solve small-scale coordination problems. Though arbitrary, on some moral conventionalist accounts, conventions would progressively acquire a normative force in the population through a psychological process of habituation. Morality would then be nothing more than a set of conventions combined with some specific subjective feelings that people have toward them (Sugden 2005). According to Binmore, fairness norms are actually 'mere' conventions solving coordination problems. What makes these conventions moral is the nature of the coordination problems they are designed to solve. The choice of resource allocation in the game of life is the kind of coordination problems that falls in the realm of morality. Thus, the conventions established in the game of morals on the basis of the OP algorithm are moral in this sense.

Two general kinds of objections can be made against moral conventionalism, one empirical, the other philosophical. I do not regard any of them sufficiently conclusive, which makes moral conventionalism an attractive meta-ethical stance against moral skepticism. To start with the empirical objection: moral conventionalism goes against a significant body of literature in empirical moral psychology which demonstrates (via experiments) that moral and conventional judgments differ in nature. Conventional judgments depend on conventions and thus respond to social practices, are arbitrary, and historically contingent. By contrast, moral judgments are generally regarded as lacking these three properties. Relatedly, moral and conventional rules are grounded on different kinds of judgments, and the conventionalist claims that morality is conventional is thus empirically false.[14] In particular, the empirical literature seems to establish that children of three years old, faced with some specifically designed tasks, exhibit an ability to distinguish between moral and conventional rules. Philosophical accounts interpreting these experimental results locate the distinction between morality and conventions both in *form* and in *content* (Southwood 2011). Regarding their form, moral rules tend to be characterized as non-contingent and global in scope, while conventional rules are characterized as contingent and more local. Regarding their content, it is suggested that moral rules deal with essentially other-regarding as well as impartial behavior and issues related to welfare, harm, fairness and trust. Conventional rules instead regulate self-interested behavior in the context of agreed-upon social practices.[15] Ultimately, it has been argued that the empirical evidence supports a conception of morality which has four constitutive properties—these are: seriousness, generality, authority-independence and objectivity. Conventional judgments and rules are believed not to have these properties (Kumar 2015).

It could be argued that the empirical evidence and its philosophical interpretations cast doubt over the relevance of moral conventionalism. It might be argued that the salience of the moral/conventional distinction for children or even adults is due to moral naiveté or

[14] The literature in development and moral psychology on the moral/conventional distinction is relatively abundant. The work of Elliot Turiel (1983) is generally regarded as seminal. Other important references are Smetana (1993) and Nucci (2001).

[15] Southwood (2011) argues that a philosophically more convincing way to ground the morality/convention distinction, still compatible with the empirical evidence, is by reference to whether or not a rule is practice-dependent. Specifically, contrary to conventional judgments, moral judgments are claimed to be practice-independent, i.e. they do not depend on the existence of a socially-agreed upon practice in the relevant community.

cognitive error. But, as a proponent of a Humean account of justice and morality would recognize, the evidence in support of that the distinction appears to be ignored *only* by persons with psychopathological tendencies "gives pause for thought" (Sugden 2008, 3). Which is to say, "It would be disturbing to have to conclude that psychopaths have a better understanding of the nature of morality than psychologically normal people do." (Sugden 2008, 3). Still, I think that the moral conventionalists can answer the empirical challenge of moral psychology in a way similar to Sugden (2008), who suggests that there are reasons to think that the very moral/conventional distinction is itself conventional. First, it should be noted that the empirical evidence is not as straightforward as some moral and development psychologists suggest. For instance, while gratuitous physical aggressions are virtually universally perceived as morally wrong, which actions belong to the category 'gratuitous physical aggressions' seem to vary across contexts and cultures (Haidt, Koller, and Dias 1993). In other words, moral judgments seem to be conventional after all. Other empirical studies establish that the transgression of some rules that are regarded as conventional (for example, rules of politeness, etiquette, and respect) in the western world, is considered to be harmful and serious in other cultural contexts (Sugden 2008). A second and related reason to doubt the empirical relevance of the moral/convention distinction is that the importance given to concepts of welfare, fairness, and trust, and which are supposed to be the objects of moral judgments is itself constitutive of western philosophy and liberal societies. As Sugden (2008: 20) notes, even proponents of the distinction tend to recognize that the concepts of welfare, fairness and trust should be understood subjectively. Of course, taken seriously, such a claim would entail that the very definitions of welfare and fairness can be the subject of conventional judgments, thus ultimately undermining the moral/conventional distinction.

The second objection against moral conventionalism is more philosophical and targets another distinction, that is: How can we distinguish moral from non-moral conventions? We should be able to discriminate between conventions that 'merely' solve coordination problems without any moral significance (e.g., on which sides of the road should we drive?) and morally loaded conventions (e.g., how should we punish murder? How should wealth be distributed in a population?). Binmore's account does not offer such a distinction but we can develop some considerations on this issue. In particular, it may be argued that there are indeed authentic *moral* conventions. This is plausible even though there is a trap here: the fact that there are *moral reasons* to follow a convention does not make the convention a *moral* convention (Marmor 2009). On some accounts, I clearly have a moral reason to follow the convention about which side of the road one has to drive on, since not following it could lead to injuries or even deaths. What would be moral conventions then? Marmor suggests that the role of such conventions "is to mediate between abstract moral ideals and their concrete realization in our social interactions" (2009, 149). Consider the fact of giving to a charity. The latter is a moral ideal that gives indications and reasons for action. However, this is a very abstract ideal which leaves many issues unanswered: How much to give? To whom? How often? Marmor suggests that "[i]n such cases, conventions may evolve that specify norms of behavior that instantiate the moral principle of charity" (2009, 150). This definition is somewhat in accordance with our discussion of the moral/conventional distinction above. I have noted that while some actions, such as arbitrary physical aggressions, are universally condemned as morally wrong, the very characterization of an arbitrary physical aggression is itself conventional. It might be objected that on this account of moral conventions, conventions do not *create* morality but rather instantiate it. This seems to be quite different from the strongest forms of moral conventionalism (Binmore's included), which claim that morality is constituted by conventions. This is not really convincing, however, because the creation/instantiation distinction is actually illusory. Either we can maintain that moral properties and facts exist independently of social practices and are not created by them; in which case the moral skeptic's critique applies. Alternatively, we can maintain that moral properties and facts are practice-dependent; in which case morality is created at the same time that it is instantiated through social practices.

A stronger objection can be made, however. Indeed, the skeptic's critique can be reformulated along the following lines: why should we give any moral significance to conventions that (by definition) are ultimately arbitrary and contingent? The next section deals with this objection and argues that while moral conventionalism can eventually answer it, Binmore's specific account does not.

5. WHY BINMORE'S MORAL CONVENTIONALISM DOES NOT ANSWER THE SKEPTIC'S CRITIQUE

The skeptic's strongest objection to moral conventionalism relies on the claim that because moral conventions have amoral origins, they cannot have the kind of normative authority that any moral prescription is thought to have. Recall that one of the constitutive features of conventions is their arbitrariness. A minimal reason to follow a (moral) convention *C* is the expectation that others will also follow *C*, while there would be a reason to follow an alternative (moral) convention *C'* where one would expect others to follow *C'*. This has at least two implications. First, the reasons to follow a convention are *never* fully desire-*independent*. It depends on having appropriate preferences such that conforming to the social practice is best for the individual. Second, while I may have a desire-independent reason to follow a moral convention *C*, this reason can in principle be dominated by other desire-dependent reasons. The prisoner's dilemma is, of course, the prototypical case of such a situation. But it may also occur in pure coordination games, where while one may have a desire-independent reason to follow a moral convention *C*, the sole fact of expecting others to follow *C'* is sufficient to lead one also to follow *C'*. It follows that moral conventions do not have any necessary practical authority. As I noted in section 3, the lack of necessary practical authority is one of the skeptic's arguments against moral naturalism, and the very same argument could be used against moral conventionalism.

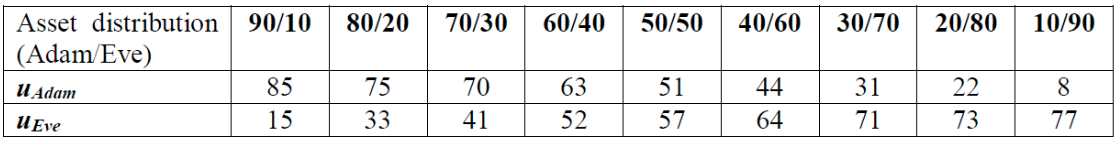
Of course, the moral conventionalist can respond in a way that is not available to the moral naturalist. The conventionalist may answer by claiming that moral conventions need not be endowed with any particular force and need not generate unconditionally dominant desire-independent reasons for action. The fact that people follow conventions to resolve issues related to morality or fairness should be taken as such, and there is nothing special about it. It might be argued that *why* people follow such conventions is irrelevant to our understanding of morality. I think Binmore, as well as other moral conventionalists, would be perfectly satisfied with this answer. Fairness norms have evolved as a coordination device in morally loaded coordination problems, and it is a fact that people follow them, which is, in itself, the evidence for the belief that they accept them. It is possible that there is nothing more to say about morality or fairness. In essence, this is very similar to Daniel Dennett's claim that moral norms function as "conversation-stoppers"

(1996, 506): they put an end to debates that otherwise cannot possibly be solved by finite computing machines.

This, however, leads to a further and ultimate difficulty. Suppose we accept all the conventionalist's claims and arguments. Together, they form a set of propositions about morality and fairness that we can denote as theory *T*. Binmore's account is a specific variant of *T,* but other similar conventionalist accounts are also instances of *T*. Suppose now, analogically to what macroeconomists are routinely assuming, that people form 'rational expectations'. By this, I mean that their beliefs and preferences about moral issues and matters of fairness are well-informed, *i.e.* they are generated on the basis of all the available and relevant information *I*. Suppose that people follow a set of moral norms and conventions *N* without necessarily ascribing to them a particular moral or normative value. Now, a critical test for conventionalism depends on the plausible answers we could give to the following question: *should people following N without knowing T continue to follow N once T is included in I?* For instance, learning that my belief that the current wealth distribution is fair is grounded on a norm that results from past bargains, where some agents had bargaining advantages (say, they were more skilled), should I continue to use this norm to form my beliefs about the fairness of the wealth distribution?

This question builds on the same intuition as Joyce's belief pills analogy, but is more about practical than theoretical reason. In essence, why should I continue to accept and act upon a particular claim or judgment about an issue once I realize that it originates from circumstances that have nothing to do with the issue at stake? It is plausible that a person introduced to Binmore's account, realizing that the fairness norms she is following result from power relations, should at least start to reflect on whether there are relevant reasons to continue to abide by the norms. Of course, since fairness norms are equilibria in the game of life and in the game of morals, the unilateral deviation is impossible (or at least irrational). By the very definition of the equilibrium concept, a player cannot increase his utility (measured according either to his personal or empathetic preferences) by using a different strategy. But, a coalition of disadvantaged individuals could in principle rationally deviate from the current equilibrium if they succeed in coordinating to change their behavior simultaneously. This would lead, in turn, to a shift in the corresponding empathetic equilibrium.[16]

I do not think this problem necessarily undermines moral conventionalism, though. However, at this point, I would like to distinguish between Binmore's naturalistic and conventionalist account of fairness and another form of moral conventionalism that, taking inspiration from Gauss (2013), I will call 'Moral Conventionalism with Public Justification'. To understand the point of the distinction, it is useful to give a numerical example to illustrate how fairness norms solve coordination problems in Binmore's account. The example will make it clear why Binmore's account is vulnerable to the above critical test. Consider two individuals bargaining in the context of the game of life over the allocation of some divisible asset. Figure 1 below gives the payoffs (expressed in terms of von Neumann-Morgenstern utilities) of the two players (that, following Binmore, I name Adam and Eve) as a function of the asset distribution. The players' utilities are arbitrarily set on a 0-100 scale, and I assume that in cases where players fail to agree over an allocation, the asset is lost and both obtain a payoff of 0:



Asset distribution (Adam/Eve)	**90/10**	**80/20**	**70/30**	**60/40**	**50/50**	**40/60**	**30/70**	**20/80**	**10/90**
u_{Adam}	85	75	70	63	51	44	31	22	8
u_{Eve}	15	33	41	52	57	64	71	73	77

Figure 1

As indicated in figure 2 below, the Nash bargaining solution **N** corresponds to the allocation where Adam obtains 60 percent of the asset and Eve 40 percent (**D** is the disagreement point). Now, suppose that two individuals, John and Oskar, have to bargain over the asset and use the prevailing fairness norms to coordinate. In Binmore's framework, that means that John and Oskar are playing the game of morals and are using the OP device to solve their coordination problem. Following Binmore, we assume that no external authority can enforce the agreement obtained behind the veil of ignorance. Both players have to assume that they have an equal chance of being Adam and Eve once the veil is removed. As indicated in section 2, it follows that Oskar and John will bargain under symmetric empathetic preferences and will

[16] This issue cannot be dealt with in Binmore's framework since all his discussion is restricted to two-person bargaining games (though two-person games can be cooperative of course).

implement the egalitarian solution. Denote as *U* and *V* the unit of the empathetic scales that both Oskar and John use to value Adam's and Eve's payoffs respectively. As shown by Binmore (1998, Chapter 4), the value of *U* and *V* can be determined by choosing them such that the egalitarian solution *with Oskar's and John's empathetic utilities* correspond to the Nash solution *with Adam's and Eve's personal utilities.* Hence, we should have 63/U = 52/V, or U ≈ 6/5 V. Arbitrarily setting *V* = 1, we get *U* ≈ 6/5. These values indicate how Oskar and John trade Adam's and Eve's personal utilities behind the veil of ignorance to reach an agreement. In this example, 6 units of Adam's personal utility are judged to be worth approximately 5 units of Eve's utility.

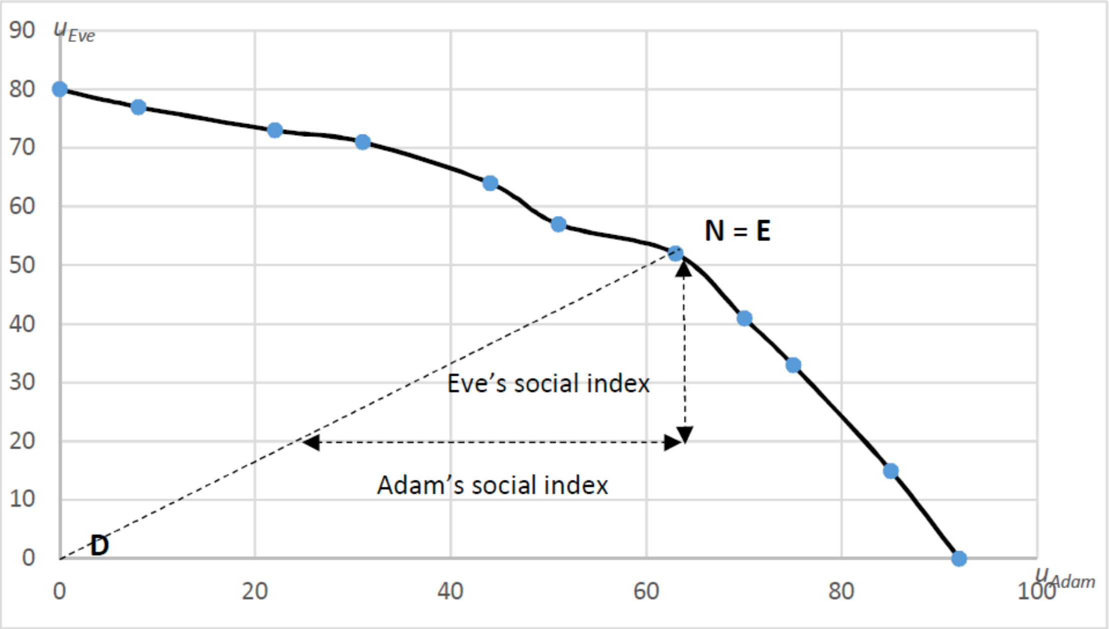


Figure 2

The empathetic preferences whose scales are determined by the variables *U* and *V* encapsulate the fairness judgments that Oskar and John use to solve coordination problems. This can be seen more clearly if we assume that the available quantity of the asset increases. As depicted in figure 3, this induces an expansion of the Pareto frontier and a modification of the Nash solution. The Nash solution now corresponds to the 50/50 bargain. However, in the short run, empathetic preferences remain unchanged by assumption. Oskar and John will thus continue to trade 6 units of Adam's personal utilities for 5 units of Eve's personal utilities. Using the OP device to coordinate, Oskar and John will implement the egalitarian solution for the *new* bargaining problem but will use their *original* empathetic preferences. This leads to the coordination somewhere between the 60/40 and the 50/50 allocations. In the short run, the Nash and egalitarian solutions will thus no longer coincide, until cultural evolution induces a modification of empathetic preferences. Over the medium and long run, fairness norms are thus determined by natural and cultural evolution, especially bargaining power. But over the short run, they are used to coordinate in bargaining problems and do not reflect *current* bargaining power.

Binmore's fairness norms clearly have all the characteristics of moral conventions: they are grounded on past and current social practices, they could have had different content if past bargains had been different, and they are arbitrary in the sense that different empathetic preferences would also permit coordination on the Pareto frontier. Moreover, though this is not made explicit in Binmore's account, individuals bargaining under the veil of ignorance expect that the agreement corresponding to the egalitarian solution, given current empathetic preferences, will be reached.[17] I would argue, however, that it is insufficient to pass the critical test presented above. To see why, consider the reasoning of Oskar who ends up being Eve once the veil of ignorance is removed.

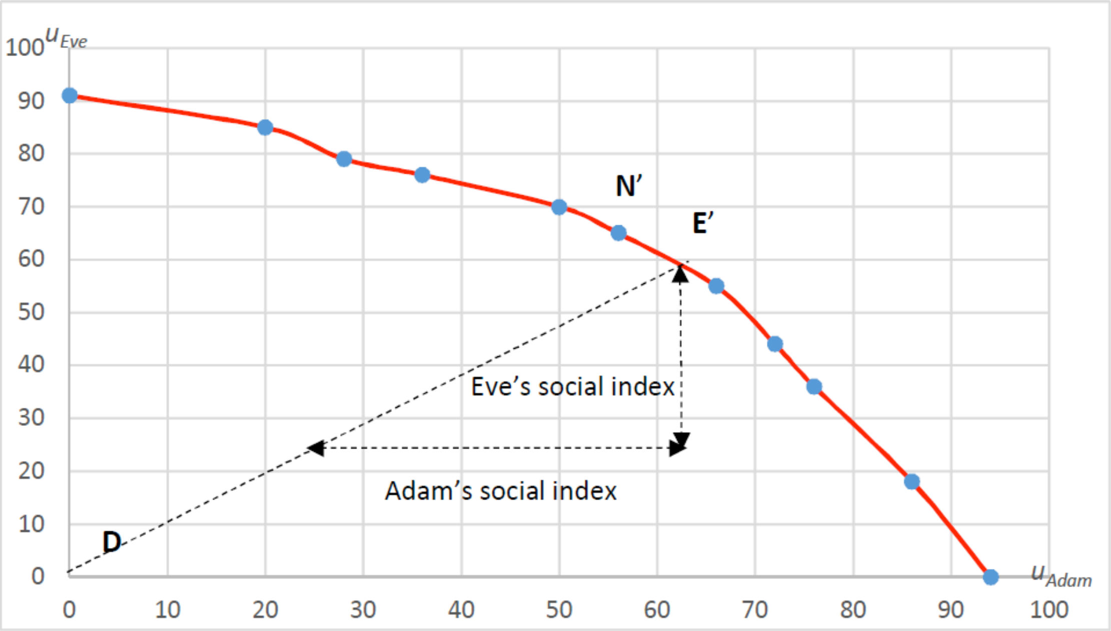


Figure 3

[17] This is true if we assume that players know their (identical) empathetic preferences and know (or at least strongly believe) that they are identical across the population.

Once the Pareto frontier has expanded, Oskar will obtain approximately 45 percent of the asset through playing the game of morals with John (whom we assume ends up being Adam once the veil is removed). Oskar is thus slightly disadvantaged, but *from his perspective*, the result is fair. This is due to the fact that the agreement is obtained by using Oskar's and John's empathetic preferences which they have inherited from past bargains. However, were Oskar to realize that his empathetic preferences are the result of bargains made in the past and whose outcomes have been determined by an *old* bargaining power structure that no longer prevails, there is no reason that to think that he would agree to an allocation that is worse for him than the one he would obtain using his *current* bargaining power. Indeed, actually, Oskar could use threats to implement the Nash solution and obtain half of the asset. John would, of course, disagree, arguing that by the prevailing standards the allocation is fair. But of course, this begs the question: Oskar would reply that what is fair is determined by bargaining relationships and that there is no reason to use *past* bargaining power rather than *current* bargaining power to allocate the asset. The point is that fairness norms play the role of coordinating devices if and only if individuals fail to reflect over the content and the origins of their empathetic preferences. Then, in this latter case, fairness norms are moral conventions that indeed play the role of Dennett's conversation stoppers. They put an end to the bargain and avoid costly negotiations.

Now, this may be an adequate account of how individuals actually solve coordination problems. It may be the case that in many situations, we play the game of morals almost automatically without reflecting on the content of our fairness judgments. The latter are just what they are, we expect others to make similar judgments, and we do not give more thought to this. However, this is not sufficient, because Binmore's account is explicitly about both descriptive *and* prescriptive ethics. While we may grant that this is an interesting account of individuals' actual reasoning in coordination problems, it is definitely not a convincing account of what makes morality special. Consider once again the above critical test. Undoubtedly, some persons in the population, even knowing theory *T*, would answer 'yes' to it. The reason for this is simply that it would be in their personal interest to continue to follow the set of norms *N* (it would probably be the case of John in our example). Cynics would concur: fairness norms are often nothing more than a '*cache-misère*' and advantaged people are well aware that fairness is a convenient justification for the preservation of an unbalanced *status quo*. In some way, this is a vindication of Binmore's claim that fairness is ultimately grounded in power and nothing more. But this also shows that a pure behaviorist view of morality is ultimately untenable. In other words, the conventionalist cannot safely ignore the motivations and the reasons for action that are underlying the norm-abiding behavior, especially in the case of moral norms. This is well put by Philip Kitcher:

> … it's important to demonstrate that the forms of behavior that accord with our sense of justice and morality can originate and be maintained under natural selection. Yet we should also be aware that the demonstration doesn't necessarily account for the superstructure of concepts and principles in terms of which we appraise those forms of behaviour (Kitcher 1999, 222-3).

In Binmore's account, the morality of fairness norms is epiphenomenal since ultimately it reduces to (rather than merely supervenes on) power relations. Conventionalists like Binmore have to argue that there is nothing more to morality than 'conversation-stopping' devices. At the same time, if once they are aware of the genealogy of their fairness judgments, people continue to abide by them only because of necessity or personal interests, this should arguably matter to any account of morality. If there is nothing distinctive about morality, one may wonder why it is worth seeking to provide it with naturalistic foundations.

However, moral conventionalism is by no means condemned to failure. As I anticipated above, 'Moral Conventionalism with Public Justification' avoids almost all the difficulties discussed in this section. This form of moral conventionalism is grounded on its endorsement of what Gauss calls the "Public Justification Principle" (2013, 80):

> *The Public Justification Principle*: If a moral convention *C* in a community *G* is endorsable by all members of *G*, *C* is a genuinely moral convention.

A *genuinely* moral convention is a convention that has the moral authority that the moral skeptic claims a moral prescription must have. The *Public Justification Principle* thus implies that there are two kinds of moral conventions: those that are genuinely moral and those that are not. The latter are moral conventions that, though they exist in the relevant community *G*, do not impose any moral obligations on the members of *G*. Whether a moral convention is genuine or not is left to the judgment of morally autonomous and competent agents. A morally autonomous and competent agent must determine for each existing moral convention if it provides justified desire-independent reasons for conforming *beyond* the desire-dependent reason constitutive of any convention. If this is the case, the agent will endorse it, i.e. the agent will follow it as long as a sufficient number of other individuals also follow it. If there are no such desire-independent reasons for following the convention, then one may be justified (though perhaps not required) in choosing not to follow the convention (for instance in the case one considers that there is an overriding desire-independent reason not to follow it). Given the fact that moral conventions solve coordination problems, there must exist, in a given community, a relative convergence over which conventions are judged to be genuinely moral. Too important a disagreement would entail that few, if any, moral conventions were consistently followed. Since a moral convention cannot exist if an insufficient number of individuals are ready to endorse it, the community would be deprived of any consistent and stable system of morality. Gauss (2013) suggests that such moral stability and consistency necessitate what Rawls (2005) has called ‘public justification’: there must be some public knowledge of which moral conventions are endorsable by all the members of the community. Conventional morality cannot exist without such public justification.

I think that *Moral Conventionalism with Public Justification* succeeds in passing the critical test. It also helps to make clear why Binmore’s account fails: in Binmore’s account, nothing indicates that empathetic preferences are publicly endorsable. This failure is due to the fact that what makes conventions genuinely moral is their ability to be endorsed for reasons that all members of the relevant community would accept after careful moral reflection.

6. CONCLUSION

This article has provided an examination of Binmore’s game-theoretic account of fairness as an instance of moral conventionalism. I have suggested that Binmore’s naturalization of the OP leads to a view that morality is conventional. In this sense, it seems to provide an answer to moral skepticism. However, in the specific case of Binmore’s account of fairness, the moral conventionalist answer leads to a purely behaviorist view of morality and fairness. Moral motivations and reasons are then completely ignored. There is, then, nothing special in morality. Still, I have suggested that other forms of moral conventionalism that emphasize the importance of reasons that establish moral conventions escape most of the difficulties of Binmore's account.

REFERENCES

Alexander, J. McKenzie. 2010. *The Structural Evolution of Morality*. Reissue. Cambridge, UK: Cambridge University Press.

Binmore, Ken. 1987. "Modeling Rational Players: Part I." *Economics and Philosophy* 3 (2): 179–214.

Binmore, Ken. 1994. *Game Theory and the Social Contract: Playing Fair*. Cambridge, MA: MIT Press.

Binmore, Ken. 1998. *Just Playing: Game Theory and the Social Contract*. Cambridge, MA: MIT Press.

Binmore, Ken. 2005. *Natural Justice*. New York, NY: Oxford University Press.

Bowles, Samuel, and Herbert Gintis. 2011. *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton, NJ: Princeton University Press.

Dawkins, Richard. 1989. *The Selfish Gene*. Revised edition. New York, NY: Oxford University Press.

Dennett, Daniel C. 1995. *Darwin's Dangerous Idea: Evolution and the Meaning of Life*. New York, NY: Simon and Schuster.

Gauss, Gerald. 2013. "Why the Conventionalist Needs the Social Contract (and Vice Versa)." *Rationality, Markets and Morals* 4: 71-87.

Gintis, Herbert. 2006. "Behavioral Ethics Meets Natural Justice." *Politics, Philosophy & Economics* 5 (1): 5–32.

Haidt, J., S. H. Koller, and M. G. Dias. 1993. "Affect, Culture, and Morality, or Is It Wrong to Eat Your Dog?" *Journal of Personality and Social* Psychology 65 (4): 613–28.

Harsanyi, John C. 1953. "Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking." *Journal of Political Economy* 61, 434-435.

Kitcher, Philip. 1999. "Games Social Animals Play: Commentary on Brian Skyrms's Evolution of the Social Contract." *Philosophy and Phenomenological Research* 59 (1): 221–28.

Kumar, Victor. 2015. "Moral Judgment as a Natural Kind." *Philosophical Studies* 172 (11): 2887–2910.

Mackie, John Leslie. 1977. *Ethics: Inventing Right and Wrong*. New York, NY: Penguin.

Marmor, Andrei. 2009. *Social Conventions: From Language to Law*. Princeton, NJ: Princeton University Press.

Nucci, Larry P. 2001. *Education in the Moral Domain*. Cambridge, UK: Cambridge University Press.

Rawls, John. 2005. *Political Liberalism*. New York, NY: Columbia University Press.

Rubinstein, Ariel. 1982. "Perfect Equilibrium in a Bargaining Model." *Econometrica* 50 (1): 97–109.

Sinnott-Armstrong, Walter. 2015. "Moral Skepticism." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2015. Retrieved from <http://plato.stanford.edu/archives/fall2015/entries/skepticism-moral/>.

Skyrms, Brian. 1996. *Evolution of the Social Contract*. New York, NY: Cambridge University Press.

Skyrms, Brian. 2004. *The Stag Hunt and the Evolution of Social Structure*. Cambridge, UK: Cambridge University Press.

Smetana, Judith. 1993. "Understanding of Social Rules." In *The Child as Psychologist: An Introduction to the Development of Social Cognition* edited by Mark Bennett, 111-141. New York, NY: Harvester Wheatsheaf.

Southwood, Nicholas. 2011. "The Moral/Conventional Distinction." *Mind* 120 (479): 761–802.

Sugden, Robert. 2001. "Ken Binmore's Evolutionary Social Theory." *The Economic Journal* 111 (469): 213–243.

Sugden, Robert. 2005. *The Economics of Rights, Cooperation and Welfare*. 2nd ed. Palgrave Macmillan.

Sugden, Robert. 2008. "Is There a Distinction between Morality and Convention?" *Working Paper Series*, University of East Anglia, Centre for Behavioural and Experimental Social Science (CBESS) 8–1. Norwich, UK: School of Economics, University of East Anglia.

Turiel, Elliot. 1983. *The Development of Social Knowledge: Morality and Convention.* Cambridge, UK: Cambridge University Press.

Verbeek, Bruno. 2008. "Conventions and Moral Norms: The Legacy of Lewis." *Topoi* 27 (1–2): 73–86.

Young, H. Peyton. 2001. *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions.* Princeton, NJ: Princeton University Press.

Cyril Hédoin is full professor of economics at the University of Reims Champagne-Ardenne (France). His academic work essentially belongs to the philosophy of economics and to institutional economics. He has recently published papers in *Economics and Philosophy*, *the Journal of Economic Methodology* and the *Journal of Institutional Economics*.
Contact e-mail: <cyril.hedoin@univ-reims.fr>

Erasmus Journal for Philosophy and Economics,
Volume 11, Issue 1,
Spring 2018, pp. 80-99.
https://doi.org/10.23941/ejpe.v11i1.356

Philosophy With Feet in the Mud: An Interview With Ingrid Robeyns

Ingrid Robeyns (Leuven, Belgium, 1972) is a philosopher at Utrecht University, where she has held the Chair in Ethics of Institutions at the Ethics Institute since 2014. She also serves as the president-elect of the Human Development and Capability Association. Before coming to Utrecht University, Robeyns held the chair of Practical Philosophy at the Faculty of Philosophy of Erasmus University from 2008 till 2014. She received an MSc in Economics from KU Leuven in 1997, an MA in Philosophy from the Open University in 2007, and a PhD from the Faculty of Politics and Economics from Cambridge University in 2003. Her PhD thesis, on applying the capability approach to gender inequality, was supervised by Professor Amartya Sen.

Robeyns does research in analytical normative philosophy, in particular on theories of justice and applied questions. She also regularly conducts interdisciplinary research. Robeyns is the author of *Wellbeing, Freedom, and Social Justice: The Capability Approach Re-Examined* (2017a). Her work has appeared in various journals, including *Ethics*, the *Journal of Economic Methodology*, the *Journal of Human Development and Capabilities*, and the *Journal of Political Philosophy*. Robeyns is the principal investigator of *Fair Limits*, a research project funded by a European Research Council (ERC) consolidator grant, on the question whether there should be upper limits on the amount of financial and ecological resources a person can have. She was elected a member of the Royal Netherlands Academy of Arts and Sciences this year.

The Erasmus Journal for Philosophy and Economics (EJPE) interviewed Robeyns about her formative years, her scholarship on the capability approach, the Fair Limits project, the relevance of political philosophy for public policy, and her advice for young philosophers aspiring to an academic career.

EJPE'S NOTE: this interview was conducted by Huub Brouwer and Manuel Buitenhuis. Brouwer is a PhD candidate at Tilburg University, specializing in desertist and luck egalitarian theories of distributive justice. Buitenhuis is policy-advisor at the directorate for the Labor Market and Social-Economic Affairs at the Dutch Ministry of Social Affairs and Employment. He conducted this interview in a personal capacity. Thanks to Annikka Lemmens for transcribing the interview, and to Willem van der Deijl, Bart Engelen, and Julien Kloeg for discussions on Professor Robeyns's work.

EJPE: Professor Robeyns, you started out your academic career studying economics at Leuven. What motivated you to do economics?
INGRID ROBEYNS: I think the essence is that I wanted to contribute to improving the world. Not necessarily directly, because otherwise I would have gone into politics, or so. But I thought economics was a good subject to contribute to improving the world, because in the end economists rule the world.

In what sense do you believe economists rule the world?
Well, I think it's actually not true anymore. It’s more and more business people who rule the world, if any group can be singled out. But when you think about actors that may have the power to improve the world, it's in the first place the government and governmental institutions. And I think the government is mainly made up of economists and lawyers. But law never attracted me—I don't know why. Probably because I don't like learning by heart!

Was it clear to you from the outset that you wanted to become an academic?
No, no, absolutely not. My grandmother's sister, who was as a second grandmother to me, always said that I would one day be the prime minister of Belgium. So I probably had this drive to be a leader of sorts, to try to really serve the group for which I was working. But I should also say that I was talked into doing a PhD. That’s actually very important to stress, because now there are people who look at me and think ‘Oh but she really knew what she wanted, and she went for it’. But that is not entirely true.

You got talked into doing a PhD? How did that happen?
When I finished my economics degree, I had serious issues with economics as a discipline. I felt that the economics curriculum—the way I see it, and that’s an important qualifier—was too ideological. It was too detached from the world, leaving too little room for normative questions. I recall a conversation when I was part of a student group that was advocating fair prices for banana farmers in Latin America. My professor of development economics recognized me as a member of that group, and asked me why we were advocating ‘fair prices’. He said: ‘there is no such a thing as a fair price. Prices are determined by supply and demand on the market’. At the time, I did not have the tools to

analyse why I was so bothered with such statements, but I had strong intuitions that this could not be the end of the story. Those matters made me believe I should go and do something else.

I left for a year to do a rural development programme in Göttingen, but quit early because it wasn't very good. I decided to stay in Göttingen and fill the year with courses I picked from social and political sciences, and enjoyed that a lot. But I still had no clue of what to do next. I was contemplating to take another degree in development studies, political science, or philosophy. Then, some day, I ran into Erik Schokkaert, my former professor of welfare economics from Leuven. He was looking for a PhD candidate, and he really talked me into taking that position. I remember that that conversation was, in retrospect, actually quite embarrassing for me. I essentially said to Erik that I couldn't do a PhD in economics, because it's a right-wing science. He was trying to convince me by saying that I could work on inequality, and on poverty and gender issues. In retrospect, if I were him, I would probably have thought: 'well if she doesn't want to do it, then fine—I'll find someone else'. But he supported me, and without his encouragement I am not sure I would be in academia today.

Were there other people that were of particular influence to you?

Amartya Sen is the obvious one. Erik Schokkaert and I agreed that it would be good for me to go abroad for a year. Erik suggested that I could work with someone like Stephen Jenkins, who does empirical poverty analysis. But then I was in the pub with the sociologist Sarah Bracke late at night and she asked me who I would like to work with, if I were to have a totally free choice. I said that it would be great to work with Amartya Sen, but that this was impossible. When Sarah asked why it was impossible, I responded 'listen, he's this big guy, how could I go and work with him?'. Sarah then made me promise that I would write to Sen, and every time I would see her, she would ask me 'Have you written to him?'. In the end I did write to Sen. I was of course nervous about that. I wrote that I was working on a PhD dissertation on gender inequality, using his capability approach, and asked whether I could spend a year with him. He said yes. That's how I came to Cambridge.

On a sidenote, this may actually be seen as a case of adaptive preferences. We always think adaptive preferences are for poor and oppressed people, but we also suffer from thinking that things are impossible, when actually they are possible. We are simply socialized in believing they are not.

What I discovered at Cambridge was a whole new world. It's not just the level of the people there, but also the ease with which you could do interdisciplinary work, especially compared to the academic system in Belgium. In Cambridge, which has colleges, I met all these people from philosophy, sociology, history and many other fields. I really liked that. The intellectual freedom was amazing. So I didn't want to go back to Belgium. I applied to stay in Cambridge and Sen became my supervisor.

At Cambridge, I also interacted with many other scholars, such as the feminist theorist Juliet Mitchel, for whom I taught a course on gender inequalities, and the economic historian Jane Humphries, who introduced me to the world of feminist economics, which was very important for me at the time. Jane and I also co-edited, with Bina Agarwal, a double special issue of the journal *Feminist Economics* (2003). We did this while I was a PhD student, and I learnt a lot from her about the social rules and expectations of academia.

Knowing what you know now, would you still have studied economics? Or do you think you would have studied philosophy instead?

Well, in the end I studied both, but at first I started with economics. I'm very happy that I studied economics, because it makes you immune for thinking that money falls out of heaven, which some philosophers suffer from. It also prevents you from having these overly simplistic assumptions, like some radical egalitarians, who, in my view, do not take feasibility constraints and incentive objections sufficiently seriously. As an economist, you're always trying to think about efficiency; as a political philosopher, you're trained to always think of distributive consequences. The nice thing about studying both disciplines is that you never forget either.

The other thing that I'm really grateful for, is that I know how to read statistics. If you want to say something about the world as it is, you have to be able to do that. There are some philosophy programmes where philosophy students have to do a minor in another discipline, like psychology or biology. I think that's very good. I actually think it's better not to study *only* philosophy.

What I really like doing, in the end, is to try to come to all-things-considered judgements. This happens a great deal in applied ethics, but not always in political philosophy. To make all-things-considered judgements, one always has to include some empirical information in the analysis. Hence it is an important skill for an applied political philosopher to know how to judge the quality of empirical research and to be able to read and interpret quantitative data.

Let's move to your recently published book: *Wellbeing, Freedom and Social Justice: The Capability Approach Re-examined* (2017a). We're interested in this last word, 're-examined'. Why was a re-examination necessary?

The capability literature is relatively young if you compare it to, let's say, utilitarianism. In such a young literature, after a while, somebody needs to weed out mistakes and clarify stuff. Much of my work on the capability approach has been to clarify things and to try to bring structure to the discussion. What I tried to do in that book is to provide the most general account of the capability approach that is possible.

Writing such a general account is important, because there are many different people working within the capability approach. Sen and Nussbaum are most famous, but there are many others. This then prompts the question: What unites all that different research? The variety of research that claims to be within the capability literature is huge—from capabilitarian theories of justice, to inequality measurement, to discussions about curriculum design in schools, to social policy proposals for welfare states. Is it really coherent to say that all this research shares a common core? If so, what is that core?

It can be helpful to see that there is a distinction between the capability approach and capability theories. The capability approach is the general thing, and capability theories are the particular instances of the approach where you fill in the details. Sen, for example, has developed the outlines of the capability approach—but it's still very sketchy. Nussbaum has developed a capability theory.

In addition to generalizing, I wrote the book to help people in different disciplines to understand each other better. Finally, I raise a range of questions that people that are new to the literature pose again and again. You could say this is the 'frequently asked questions' part.

Do you think measurability is a problem for the capability approach?

We know it's very hard to measure freedoms. Take surveys, for example. There is a limit to how long a survey can be, because the response rate drops if the survey is too long and you'll get more sample selection bias.

The problem with the existing empirical studies in the capability literature is that most data are, at best, proxies for functionings and capabilities. Sometimes there aren't even clear proxies available. How do you measure stress, for example, let alone the freedom not to experience excessive amounts of stress? And how do you go about measuring the different combinations of functionings that people can realize?

Then there also is a policy problem: the capability approach is generally insufficiently specified to make concrete policy proposals. This is because making policy is not only about functionings and capabilities, but also about one's views on the appropriate division between societal and personal responsibility, the appropriate weighing of issues of fairness and sufficiency, and many other things.

I think the most powerful contribution of the capability approach is to show that if you have a purely money-based policy framework, then you're missing out on important dimensions. An example is the debate surrounding government incentives for women on the labor market, in which an increase in the number of women who have paid work is seen as a good thing. Such an increase isn't necessarily good, however. Whether it is, depends on whether women wanted this, what the costs are, what the distribution of care work within the household is, and so on. If you only look at economic indicators, you will miss out on these things.

Do we really need the capability approach to tell us that? Most economists would agree that money doesn't cover everything.

Yes, that's true—in theory. What I found is that in economics, and somewhat less in economic policy, there is this huge gap between what's possible in theory and what happens in empirical work. In theory, economics works with utility. But then the question is, what is utility? If you look at how it's translated to empirical work, there are these assumptions that make you end up analysing disposable income, or purchasing power, or comparable metrics. The econometrician Wiebke Kuklys and I wrote a paper together in which we show how a set of assumptions lead to this jump from welfare to income metrics, and how problematic the underlying assumptions are. That paper was ultimately published as a chapter of her book that was published a few months after her tragic early death in 2005.

There is something interesting going on in those discussions. I notice that if I talk with economists about the contribution of the capability approach or about other criticisms of economics, they are almost always very defensive. Or they will come up with examples where they actually did something that could counter the capability critique. But in the totality of the literature, these counter-examples are rare, and I believe that the critique of the capability approach on mainstream economics remains valid.

Your point is clear when it comes to the academic arena. Economists have a way of doing economics in which the domain of money is larger than seems justifiable. But if we look at the domains of policy and politics—civil servants, politicians—it would seem that they do weigh different domains. And the domain of money is smaller than in academic economics.

You're right that if my criticism has a bite, it's probably first and foremost for the academic world. But then it must be said that the academic world, of course, has an effect on policy-making. Take, for example, the number of people who live in absolute poverty in the world. That's determined by the absolute poverty measure of the World Bank—I'm not absolutely sure what it is now, but I think it's about 2 dollars a day. That's a money-based metric, and there are all these studies from development economists that show that we understate the incidence of poverty that way. This flows over into politics: on the basis of this monetary metric, some people say that globalization leads to a decrease in the total amount of people living in poverty. I'm not saying that's not true, but only that these are political consequences from the use of academic research, in this case the income-based measures.

Why are economists so slow on the uptake regarding these criticisms?

To use a bit of an exaggeration, I think economists are socialized to believe that they don't need to really engage with other disciplines. And if they do engage in other disciplines, they do it in a way that doesn't really respect those disciplines enough. Take the example of the work on identity economics by Akerlof and Kranton (2000). To economists that was very new, but if you look at sociological work on the topic, you know this is actually just all these sociological insights captured within an economic framework. Even heterodox economists, such as the feminist economist Nancy Folbre, had done work like that at length, but that work doesn't use formal models, and hence is not acknowledged for what it actually contributes.

Economists sometimes reinvent stuff that has been done by other disciplines for a very long time. They should have a more open-minded attitude, in which they see other disciplines as genuine epistemic equals. I think there is some truth to the view, common among many non-economists, that economists often have an arrogant attitude towards other disciplines. And, of course, if you have that attitude, how can you learn from other disciplines? If you cultivate and socialize new economists in a way that makes it hard for them to learn from other disciplines, there will be quite a number of people, like me, who quit economics and move to neighbouring disciplines. There are plenty of excellent economists working in history departments, or in political science departements. They just couldn't do what they wanted to do within the economic discipline. In this way, there is a kind of disciplinary cleansing in economics: if you don't fit the quite strict methodological and paradigmatic requirements of what economics is supposed to be, you get frustrated and you leave.

And how should we change that? Do you have ideas on this?

I've stopped seeing this as my problem. You can't solve all problems, right? You have to pick your battles and I just think there is other work to be done. But I still think economics as a discipline should change. I also appreciate, however, that there is an increasing number of people who work on economic topics outside of economics. A good example is my colleague Bas van Bavel, an economic historian who works on long-term developments in capitalism and on inequality in wealth. Jane Humphries, the economist who was important for me when I was studying in Cambridge, also moved to economic history. The increasing number of PPE-programs shows a similar trend. It seems that rather than changing the house of the economists, some people are now building a new house, where economic issues can be analysed with a plurality of methodologies and ontological assumptions. I think that's a much more constructive project. What remains important is that people who study economic problems from these different viewpoints find their way into civil society, agencies such as the Dutch Central Planning Bureau, and other government structures.

Let's now move on to your ERC-funded Fair Limits project. When did you first get the idea for limitarianism?

Around 2012, it struck me that so few people were actually studying the rich, and decided to work on that topic when asked to give a keynote at at a graduate conference of the Erasmus Institute for Philosophy and Economics (EIPE). I wondered whether it was possible to construct something like a poverty line, but then for rich people. There are people who have done empirical work and who just say: The rich are the richest 1%, then you have the richer, the 0,1%, and then you have the richest, the 0,01%. It seemed to me, however, that this is an unsatisfying way of conceptualizing richness. To get to a more satisfactory conceptualization, you could use the theoretical debates on the poverty line that happened mainly during the 1970s. So I first developed the richness line and then started to think about whether there are normative issues related to what we would call 'super rich people in society'.

You have hired a research team on the Fair Limits project. Could you sketch for us what you would like the team to have achieved in 2022, the year in which the project finishes?

There are two PhD positions in the project. Petra van der Kooij is working on ecological limits. The normative case for ecological limits is not too difficult. Given the limited capacity of the atmosphere to absorb greenhouse gases, none of us has a right to pollute without limit. The interesting questions regarding ecological limits are more about the speed with which we are making the transition to less pollution, how to deal with global inequalities in pollution, and what the duties of individuals are given that government policies are moving so slowely. Dick Timmer is working on limits on economic resources. There, the normative case for a limit is more complicated, because economic resources do not fall from heaven: somebody makes them. An interesting question here is to what extent limitarianism already follows from—or, is compatible with—existing views, such as Rawlsian egalitarianism or luck egalitarianism.

Next to the PhDs, there are also two postdoc positions on the project. Colin Hickey works on philosophical methods for the world as it is, which relates to the whole 'ideal theory' versus 'non-ideal theory' debate. I decided to make the entire project committed to the non-ideal turn in political philosophy, but there are still many questions related to the methods non-ideal political philosophers should use. I hope that Colin and I can make some progress on that front. Tim Meijers works on who the agents of justice are. Political philosophy often just assumes that the state is the only agent, and we think that's problematic. One of the ways that the state may not be the sole agent, is that rich people can be encouraged to give away surplus money in philanthropy. Tim and I have written the philosophical chapter for a report of the Netherlands Scientific Council for Government Policy on philanthropy and policies towards philanthropy in the Netherlands.

The fifth project is the one that I will do myself. I will look at what we can learn from non-Western philosophies when it comes to limits on economic and ecological resources—think of Chinese philosophy, Ubuntu philosophy, and indigenous philosophy. I will also work on synthesizing the five subprojects.

In your original article on limitarianism* (2017b)*, you say that you are concerned only with "non-intrinsic limitarianism, and remain agnostic on the question of whether intrinsic limitarianism is a plausible view" (5).*[1] *Intrinsic limitarianism is the view that "being rich is intrinsically bad"; non-intrinsic limitarianism is the view that "riches are morally impermissible for a reason that refers to some other value" (5). To what extent do you think that intrinsic limitarianism could be a plausible view?

I doubt that it is a plausible view. The most obvious way you could argue for intrinsic limitarianism, is if you adopt an account of the human character or the human person in general, on which it harms you as a person if you are rich. On such a view, it is intrinsically bad to be rich, no matter what effect that has on public values. Perhaps you could defend this claim with a secular virtue-ethical account. Another possibility, if you are interested in philosophy of religion, might be to argue on the basis of the Bible or the Koran that a religiously virtuous person is a non-rich person. I haven't thought all of this through carefully, but I am doubtful that one can make a convincing argument for intrinsic limitarianism.

[1] All references from here onwards are to Robeyns (2017b), unless otherwise indicated.

The central claim of non-intrinsic limitarianism is that it is morally impermissible "to have more resources than are needed to [lead a] fully flourishing life" (2). What is a fully flourishing life?
I leave that open in my paper. I just say that we should decide this through a political process. There are colleagues who challenge me, saying that I should bite the bullet and provide a detailed and precise account of flourishing. My aim in this paper was different, however. I wanted to provide the structure of an argument for limitarianism. It does seem to me, though, that a plausible account of a flourishing life would account for the widely shared intuition that at some point, you have everything you need: An increase in riches will no longer lead to an increase in your quality of life—it will only allow you to gather more stuff.

But how about people who have expensive tastes? They might still experience increases in their quality of life for increases in income above the riches line?
Indeed, a counterexample to such an account of quality of life would probably be someone who has expensive tastes, for instance someone who really has a passion for collecting art and wants to buy all Van Goghs and Gauguins that are put up for sale. There is never really enough money if you want to do that. My response to this is threefold. First, if you think about how to organise society and design institutions, there will always be cases where somebody's situation has not been properly accounted for. Expensive tastes are, possibly by definition, those tastes that are statistically rare in the population. Second, the problem of expensive tastes is not only a problem for limitarianism, but for other views of distributive justice as well: They keep coming back when preferences play a central role in theorizing. Third, I try to do political philosophy for the world as it is. Some questions may be philosophically interesting, but it's sometimes almost obscene to focus so much on them. That is related to my meta-view on what philosophy should do.

We see that expensive tastes may not be a central concern for your project, but what would you say to someone who, through no fault or choice of his own, has expensive tastes that he could, given his income, satisfy? Would we need to tell this person that his income in

excess of the riches line has zero moral worth and should be taxed away?

So, suppose you are a kid and your parents have raised you to come to believe that it's normal to have caviar every day. I think the solution there is not to say that we should accommodate those expensive tastes in the flourishing account. Instead, I think the solution should be to say these preferences are morally bad and unsustainable, and that we should help the person who has them to get another set of preferences. Preferences are made, remade, and challanged all the time. Look at how our preferences regarding smoking changed drastically in recent years. I think that you will see a similar preference change regarding meat. You just see it beginning all around you now. Hence, preference formation processes occur all the time, and if someone has an expensive taste that is suboptimal from a collective point of view, there is at least a prima facie reason that we should try to change that preference, rather than accommodating the expensive taste without asking any questions about that preference.

A possible worry here is that this way of dealing with expensive tastes would be illiberal.

I think that the political and societal effects of preference formation should be a central project in political philosophy. Many of us have embraced the liberal paradigm that preferences are sacred and should not be touched. This is one point where we've been influenced by economics: whatever the consumer wants, the consumer wants; there are no moral questions to be asked. Although I am probably in essence a liberal philosopher, I do think we should look more critically at some of our basic views, including the view that if you want to show respect to people, you should not question their preferences. I think that's really a mistake.

Liberal political philosophers, following Rawls (1971) and Dworkin (1981), often attach great importance to their theories being neutral between different conceptions of the good. Do you think there would be a way of fleshing out limitarianism that is consistent with liberal neutrality?

That is a question that I should study. I don't have an answer to it yet. Some of the obvious challenges to limitarianism are challenges on the grounds of coercion, paternalism, and, indeed, neutrality. Then again, I don't think we currently have neutral institutions either. What I really want to do with limitarianism is to study it as a framework for the world as it is. And there I think matters become muddier. Although we should try to be neutral, especially when it concerns religion, I don’t think there is a way to organise society that is fully neutral in all relevant senses, and that does not come at an excessive cost in terms of sacrificing average wellbeing.

You adduce two arguments in defense of non-intrinsic limitarianism. The first is the democratic argument, according to which there should be a limit on how much money people can have, because otherwise the rich can acquire so much political influence that the “value of democracy” (6) and an “ideal of political equality” (5) would be undermined. We wonder why, exactly, the democratic argument is an argument for a **limit*****, rather than*** **equality an sich?**

The reason is that surplus money is money that you do not need for a fully flourishing life. That means that if you have surplus money, you can basically spend it on anything, without affecting your own quality of life. People who do not have surplus money, on the other hand, can only influence politics at an opportunity cost to their own flourishing. Also, I don’t think that we need full equal opportunity to political influence and power. The reason is that we are fine with people who are smarter, or who simply have many ideas about how to run a political party, having greater political influence and power. It is problematic if they have these because of morally arbitrary characteristics such as the amount of money they have.

On a related note, your democratic argument seems to suggest that political influence increases with income. Introducing a limit would then perpetuate the issue for people below that limit. Wouldn’t a progressive tax for all incomes suit the argument better?

That depends on the details. If you were to construct a progressive tax system with a marginal tax rate on income and wealth at 70%, it could still be that someone with market luck would end up with quite a lot of surplus money. By definition, they could spend that surplus money on influencing political processes without any effect on their own flourishing. That is a difference with those situated below the riches line, since if they spend their money on political processes, it comes at the opportunity cost of their personal flourishing.

Whether a certain amount of money is enough to buy greater political influence and power depends on whether there are structures and institutions in a society that are able to shield the economic domain from the political domain. The democratic argument loses its force if it is impossible for people to turn economic advantages into political ones.

This brings me to an issue that is important to stress: We are investigating whether limitarianism is a view that can be defended. It may well be that after we have evaluated all the arguments for limitarianism, we find that none of them are very appealing. I do think, however, that the second argument I put forth in defense of limitarianism, the argument from unmet urgent needs, is quite strong. Hence, I do not anticipate that, five years from now, we will have to conclude that limitarianism should go into the dustbin of ideas.

Let's move on to this second argument for limitarianism. According to this unmet urgent needs argument, imposing a limit on how much income and wealth people can have is justified if one or more of the following three conditions holds: (a) extreme global poverty, (b) (significant) local or global disadvantages, and (c) urgent collective action problems. You point out that the argument is less demanding than T.M. Scanlon's Rescue Principle and Peter Singer's utilitarianism. On your view, we should only use excess money in order to alleviate conditions (a)-(c); not all money.

It seems that this claim relies on the assumption that all income and wealth up to the level of full flourishing has the same moral value, but that this moral value suddenly drops to zero at the level of full flourishing. If that were not the case, after all, then it seems we would be justified in taking money from those to whom it has less moral value, and giving it to those for whom it has more. Did we characterize this underlying assumption correctly? And, if so, would you be willing to defend it a bit more?

It's good that you ask this question, because you're not the first to ask it. The answer is that this definitely not an assumption I make. Limitarianism is a partial view. It says something about what should happen above the limit, but it's agnostic on what happens below it. How demanding limitarianism ultimately is, depends in part on how you fill out what happens below the riches line. The problem with Peter Singer's view is that you can no longer live your own life. You become a utility machine for solving the problems of other people. What I want to do, is take that widespread criticism of Singer's view seriously. I think that everyone who does not have unmet urgent needs should contribute to solving the unmet urgent needs of others. The richer you are, the more you should contribute, and, at some point, you should basically contribute all additional money you have—that is, your surplus money. On such a view, you can still have your own plan of life.

You write that it "may turn out that certain limitarian views … boil down to an already existing distributive view, or are compatible with an existing distributive view" (38). Have your thoughts on this developed? To what extent are certain forms of limitarianism, in fact, compatible with existing distributive views, such as luck egalitarianism, prioritarianism, or sufficientism?

There are philosophers, and I think it's very good that they raise this challenge, who ask whether limitarianism already follows from many of these existing accounts. That may be something we will conclude after five years: There is no need to do any further philosophical work on this, because in the end the distinctiveness of limitarianism lies only in matters of policy design, but doesn't change the existing philosophical theories.

However, even if it were the case that, at the level of abstract philosophical theories, limitarianism is old wine in new bottles, we still need to explain and draw lessons from the fact that it finds such resonance in public debates. I think that studying this question may tell us something about the task of philosophy, and about the fact that much of philosophy still focuses on ultimate goals and not enough on policies and institutional design.

Here is an example. Why would we think it implausible that there should be a cap on how much we can receive in inheritances over our lifetime? This would be a distincty limitarian policy proposal, but one that the vast majority of the population does not endorse at present. I see it as a task for philosophers to study whether there are good arguments for such a cap in lifetime inheritance revenue, even if it is an unpopular idea.

As a sidenote—there were proto-limitarian ideas in the history of philosophy long before the post-Rawlsian theories of distributive justice started to come on stage. Together with Matthias Kramm, I'm working on a paper in which we show that there are limitarian claims all over the history of economic and political philosophy: in Plato, Aristotle, Aquinas, Locke, Marx and many others. It may be more interesting to also connect to that earlier history.

We now have some questions regarding the relation between philosophy and policy, first continuing with limitarianism. Your paper tries to show that limitarianism can work in practice by expanding on an account of riches, the power of material resources, and a cut-off point above which riches should be redistributed. Do you believe it is a philosopher's task to show that theory can work in practice?
There are at least two answers to this question. The first is that I often do stuff that philosophers do not see as philosophy. I don't care about disciplinary distinctions. If I have a question that I find interesting, I will try to answer it. If I can't do it by myself, I will try to find scholars in other disciplines who have other types of expertise and ask them to collaborate. I am now collaborating with a group of sociologists to find out what Dutch people think about limitarianism. At Utrecht University, and I think in the Netherlands more broadly, there is fortunately increasing support for this type of interdisciplinary research. The second answer is that I prefer to do non-ideal philosophy in the sense that I want it to be action-guiding for the world as it is. If that's the kind of philosophy to which you want to contribute, then it is important to engage with relevant empirical studies, to take feasability questions very seriously, and to think about the changes in policies or institutional design that would follow.

Do you think political philosophy and policy talk enough?
No they don't, and I think it would be good if political philosophers talked more to policy-makers, politicians, and politically engaged citizens. There is still some reluctance amongst philosophers to do so. This may be explained, in part, by the fact that the type of work that some political philosophers are doing is highly abstract, dealing with counterexamples to establish, say, which abstract theory of justice is right exactly. Although such debates may be philosophically interesting, they are not necessarily useful to policy-makers.

Policy-makers have their feet in the mud: They want to know how ideas can be implemented. And, of course, both academics and policy-makers have full agenda's. I also know several young philosophers who would like to reach out more often to policy-makers and others in society, but are simply exhausted after they have done their teaching, administration and the research that is expected from them. Time is an ultra-scarce resource in academic philosophy in the Netherlands these days!

I do not think that there should be a ‘one size fits all’, however. It is good that there are philosophers like Derek Parfit, and I hope others will judge that it is good to have philosophers like me who do more of this ‘philosophy with your feet in the mud’-type of work. There is the issue, though, that very abstract, almost mathematical political philosophy is awarded higher esteem, which is, in fact, quite similar to how status is awarded in economics. Because we all want to be acknowledged and respected by our fellow scholars, this may create pressure to do work in political philosophy that is situated on the border with theoretical philosophy, rather than engaging with nonideal or policy questions.

So if we then limit ourselves to the subset of practical philosophers who both want to talk to policy-makers and do the kind of work that might be useful to policy-makers, how can we make that dialog as fruitful as possible?

It may be a very mundane answer, but I think this is a matter of learning by doing. So just do it more, take the time for it. Talk to colleagues who have done it a lot, and ask them for their advice. There is, of course, the condition that you should be given the time. Academics have a basket of tasks that often does not include talking to policy-makers. So there is a tension there. But if you put aside this practical concern, I really think it's a question of listening carefully and being sufficiently open-minded and self-critical. I've never actually had the experience that it's unfruitful.

You are quite active in the public debate, talking about parental leave schemes (2015), the funding of PhDs in the Netherlands (2014), and work pressure for academics (2018). Do you think philosophers should engage in political action more often?

This is an interesting question, because you use the words ‘political action’. Do I engage in political action? I do, in the sense that I, for example, recently sent ten tweets commenting on proposals to change the income structure for disabled workers. Is that political? Yes. But it's not political action in the sense of party politics.

Sometimes I think I should be a member of a party and try to work on a better world via a political party. But if you are a political philosopher and you are a party member, everything you do will be seen through the eyes of the ideology of that party. Moreover, I have sympathies for several parties, and talk to people from many different parties. It's different from being a professor in a field like chemistry, for example, where your expertise and your politics will be seen as two clearly different worlds.

But to come back to your question and answer it more directly: I do think that philosophers should use their knowledge to intervene when lies and distorted knowledge are produced or spread in society, or when they have knowledge or ideas that can contribute to a higher-quality democratic process, or to addressing urgent societal challenges. If we have that broader understanding of 'political action', then yes, I do think more of us should engage more often in political action.

You published your recent book in open access. Why was that?

When I received the contract already quite some years ago, it was not possible to publish open access with the prestigious academic publishers. That has changed now. Back then, I had to choose between submitting it to an academic press or publishing it open access: A trade-off between the prestige and a bit of royalties, versus accessibility. Around the time I had to make that decision, I was teaching a course in South Africa, where I was also supervising a PhD student, Ina Conradie. I asked Ina what she thought of these options, and she said open access would help her much more, both as a scholar and as a teacher. We shouldn't forget that there are huge inequalities in access to books. Some of the new generations of Black students in South Africa are simply poor—so how can they afford books? Now I get emails from people all over the world saying that they read the book and that it helps them. There are even scholars from Peru who want to translate it in Spanish, which is something that they can do with this open access book, since it's published under a CC BY 4.0 Creative Common licence, which implies that no rights need to be cleared for reproduction or translation. These results are very satisfying to me. And in the end it's only fair: we are paid by taxpayers' money, so our work should be open access.

What advice would you give to graduate students aiming to pursue an academic career in political philosophy?

Do a minor in empirical social sciences! It doesn't really matter whether it's sociology, politics, or economics: You should learn how empirical research is done. I also think it makes you more modest about what you can do as a political philosopher. In many questions in political philosophy, the arguments have empirical assumptions. But philosophers who are untrained to read empirical research, are at risk of either working with hypothetical empirical claims, or else cherry-picking those studies from the empirical literature that fit their personal views best.

Also, if you want to arrive at all-things-considered judgments, you have to find out which reasons or objections are most powerful, and that may involve reading up on empirical studies. For example, there is quite a large literature in political philosophy on basic income—the institutional proposal that every citizen should receive a regular unconditional income, independent of willingness to work or any other criterion. But several empirical scholars have argued that there is a trilemma: either the level of basic income is below the poverty line, or funding the basic income is financially unsustainable, or the basic income cannot be fully universal or unconditional. That is where the action is at this point in time in this literature. If you are a philosopher interested in basic income and can't engage with those studies, then you are relegating yourself to the margins of those debates.

A second word of advice is for PhD students who would like to stay in academia, no matter what. I'd like to tell them that if philosophy doesn't work out, there are always other options. We tend to believe that if you do a PhD, there's one route: only an academic job would make you happy. And that's really not true. I know an example of someone who started working for a Ministry after her postdoc, and initially resented that. She had hoped to find a job in academia, but it didn't work out. After two months working at the Ministry, she said she would have left academia much earlier if she had known how much fun it actually was.

REFERENCES

Agarwal, Bina, Jane Humphries, and Ingrid Robeyns. eds. 2003. "Amartya Sen's Work and Ideas: A Gender Perspective." Special issue of *Feminist Economics* 9 (2-3).

Akerlof, George A., and Rachel E. Kranton. 2000. "Economics and Identity." *The Quarterly Journal of Economics* 115 (3): 715-753.

Dworkin, Ronald. 1981. "What is Equality? Part 2: Equality of Resources." *Philosophy and Public Affairs* 10 (4): 283-345.

Kuklys, Wiebke, and Ingrid Robeyns. 2005. "Sen's Capability Approach to Welfare Economics." In *Amartya Sen's Capability Approach*, edited by Wiebke Kuklys, 9-29. Dordrecht: Springer.

Meijers, Tim, and Ingrid Robeyns. Forthcoming. "Filantropie: Conceptuele en Ethische Reflecties." WWR-Verkenning Filantropie.

Pierik, Roland, and Ingrid Robeyns. 2014. "Promoveren is Geen Werk Maar Studie." *NRC Handelsblad*, August 28, 2014. <https://www.nrc.nl/nieuws/2014/08/28/promoveren-is-geen-werk-maar-studie-1412754-a1203027>

Rawls, John. 1971. *A Theory of Justice.* Cambridge, MA: Harvard University Press.

Robeyns, Ingrid. 2015. "Respect Voor Vaderschap: Geen Woorden, Maar Daden." *Bij Nader Inzien*, June 21, 2015. <https://bijnaderinzien.org/2015/06/21/respect-voor-vaderschap-geen-woorden-maar-daden/>.

Robeyns, Ingrid. 2017a. *Wellbeing, Freedom, and Social Justice: The Capability Approach Re-Examined.* Cambridge: Open Book Publishers. <https://www.openbookpublishers.com/product/682>.

Robeyns, Ingrid. 2017b. "Having Too Much." In *Wealth NOMOS LVI*, edited by Jack Knight & Melissa Schwartzberg, 1-44. New York: NYU Press.

Robeyns, Ingrid. 2018. "Een 'Witte Staking'." *ScienceGuide*, February 14, 2018. < https://www.scienceguide.nl/2018/02/een-witte-staking/>

Ingrid Robeyns's website:
<http://www.ingridrobeyns.info/>

Fair Limits project website:
<https://fairlimits.nl/>

Erasmus Journal for Philosophy and Economics,
Volume 11, Issue 1,
Spring 2018, pp. 100-106.
https://doi.org/10.23941/ejpe.v11i1.357

Review of Anna Alexandrova's *A Philosophy For the Science of Well-Being*. New York: Oxford University Press, 2017, 196 pp.

MATS INGELSTRÖM
University of Stockholm

In *A Philosophy for the Science of Well-Being*, Anna Alexandrova urges us to revise the way we theorise about well-being. The traditional approach in philosophy is to search for the universal and most general theory of what well-being is. Alexandrova argues that this approach, to a large extent, is irrelevant and unhelpful. For most people concerned with well-being—either as policy makers trying to decide what to do, or as scientists trying to understand and measure well-being in more specific groups—the traditional approach won't do. Instead, she argues, we need to theorise about well-being in new ways.

This is a wide-ranging book with a refreshingly ambitious agenda. In it, Alexandrova consolidates the positions and arguments that she has developed and published over the recent years. The book concerns the role scientists and their scientific inquiries can and should have in our pursuit of understanding, identifying and measuring human well-being. Alexandrova wants to give an answer to "how science should define well-being, how it should measure it, and the role of philosophy in all this" (*xv*). In doing so, she explains and takes seriously recent developments in both the philosophical and the scientific field. She discusses issues ranging from theory building and concept formation, to validation and measurement. As far as I am aware, the book is unique in this way. I highly recommend anyone working in this area to read it.

The book starts with a helpful and pertinent introduction. We then get the two main parts: *Tools for Philosophy* and *Tools for Science*. Each part consists of three chapters. In two brief appendices, readers unfamiliar with the landscape are offered quick summaries of the status in contemporary work on well-being in philosophy and science. As a reader of paper books, I also appreciated the useful and detailed index.

A Philosophy for the Science of Well-Being engages from page one. Rather than making some sweeping general remarks, this review will focus on giving a rather detailed discussion of part one. Part one can be read as a unified proposal, and it is here Alexandrova advances her revisionist account of how philosophers should theorise about well-being. Part two is also interesting and contains an important collection of ideas concerning the limits of scientific inquiry regarding well-being, but these ideas are less unified and each chapter in part two would require its own detailed discussion.

In part one, Alexandrova challenges traditional philosophical views on what well-being is and how it should be investigated. Unimpressed with the traditional approach, where philosophers attempt to formulate what constitutes and grounds well-being in its most general sense, Alexandrova instead formulates and defends a position she labels *Well-Being Variantism*.

Variantism involves two claims: *Concept diversity* and *Theory diversity*. *Concept diversity* tells us that "'well-being' (and its cognates) can invoke either general or contextual concepts of well-being depending on context" (43). The context Alexandrova refers to is that of the evaluator who invokes the well-being concept, for example a scientist who seeks to characterise a well-being construct. The other part of Variantism, *Theory diversity*, makes the meta-substantive claim that "[n]o single substantive theory specifies the realisers of every concept of well-being" (43).

Alexandrova takes the implication of her Variantism to run deep. Denying the traditional invariantist position is not merely of scholarly curiosity. Alexandrova wants to significantly broaden the scope of philosophical well-being theory. If Variantism is right, the traditional search for a single unified substantive theory of well-being, exemplified in the debates between versions of the "big three" (hedonism, desire theories and objective list theories, see, for instance, Crisp 2017), at best addresses merely one of many relevant notions of well-being. So, should we believe in Variantism?

Chapter one defends *Concept diversity*. Let us look closer at what it says. First, it states that there is more than one kind of well-being evaluation. Call this part *Diversity.* Secondly, it states that the diversity depends on the context of the evaluator. Call this *Dependence*. *Diversity* indeed seems plausible. I am less convinced by *Dependence*, especially if it is understood in a deeper, fundamental sense.

Diversity gains support from observations of how 'well-being' is used. Early in chapter one, Alexandrova shows that the term 'well-being' is often used in a merely *some-things-considered* sense, in contrast to the general *all-things-considered* sense which philosophers traditionally assume. In some contexts, such as that of a sincere conversation between close friends, well-being invokes one type of evaluation. In other contexts, such as that of a policy researcher, it might invoke other evaluations. Alexandrova provides some concrete and convincing examples to support this observation.

Perhaps some would want to disqualify using ‘well-being’ for anything but general all-things-considered evaluations. If anyone holds that view, Alexandrova makes a convincing case against them (she labels this position *Circumscription*, 8-10). She argues forcefully that philosophers should take the linguistic practice of non-philosophers seriously. There seems to be no good justification, she emphasizes, for the claim that scientists and others who use 'well-being’ in merely some-things-considered senses are not *really* talking about well-being. Instead, we should accept that well-being-talk can invoke either general or partial evaluations. Furthermore, it seems plausible that which type of evaluation we are interested in may change with the context of the inquiry.

Next, turn to the *Dependence* part of *Concept diversity*. It is a bit unclear how we should interpret this dependence. Inspired by recent discussions on contextualism with regards to knowledge, Alexandrova suggests and discusses two options that would make the concept of well-being in some way depend on context. According to the *The Different Realisation view,* the threshold for when someone counts as doing well depends on context. According to *Contextualism*, the content of the concept depends on context (see 10-14). But perhaps these two are not the only possible options?

Let me propose a different explanation of what is going on here. One could point out that the general all-things-considered evaluation philosophers normally have in mind is merely one, an important but very thin, precisification of our fuzzy pre-theoretical well-being concept. There are many other precisifications which are thicker in descriptive content and thus less general, such as “physical well-being for elderly”, “student emotional well-being in elite universities” and “economic well-being of young parents”. The fact that scientists invoke different such precisifications could be explained without making the concept dependent on context in a deeper sense. In different contexts, we often have different aims and problems we are trying to solve. Our aims and problems can call for different kinds of well-being evaluations, not because the context dictates the content or the threshold of the evaluation, but because the context makes different kinds of well-being evaluations more or less relevant (or salient). A general ambition within the sciences (as well as in ordinary talk) to be relevant would then explain the diversity. This would not, however, establish a deeper dependence between context and concept. If this explanation works, we could circumvent the *Dependence* part of *Concept diversity*, while leaving *Diversity* intact. As far as I can see, there would be no harm to the overall project of the book of doing so. When Alexandrova in her afterword sums up her position, she indeed leaves out the dependence claim: "My hypothesis is that the content of the concept of well-being, at least partly, varies with context." (153).

Chapter two focuses on defending *Theory diversity*, the second part of Variantism. *Theory diversity* denies that there is a single substantive theory of well-being that covers all situations in which we make well-being judgments. Rather, what constitutes well-being may vary with context. Alexandrova structures her argument in five premises:[1]

> *Premise 1*: The philosophical toolbox of the sciences of well-being includes many, not only one, of the current theories of well-being.
>
> *Premise 2:* Depending on the context, different contents of the toolbox play a role in different constructs of well-being.
>
> *Premise 3*: Constructs of well-being, at least sometimes, specify the constituents, rather than mere causes or correlates, of well-being.
>
> *Premise 4:* Constructs of well-being in the sciences, at least sometimes, do a good job picking out well-being in a given context.
>
> *Conclusion:* So different states, as specified by different theories, constitute well-being in different contexts. (45–46)

There are some things that should be noted here. A minor observation is that one could, strictly speaking, accept the premises while denying the conclusion. An Invariantist could agree with each step, but simply claim that the only times both premise 3 and 4 hold, are when a well-being construct aligns with the single correct substantial well-being theory. This might not be the intended reading of the argument, but I fail to see an easy rewording that would close this

[1] The *toolbox*-view of scientific theories (Cartwright et al. 1995) holds that scientific theories should be understood as useful, yet incomplete, tools for constructing models that correctly represent the world (35-40).

escape route for the Invariantist. That said, the argument might still provide some inductive reason for accepting *Theory diversity*.

How plausible are the premises? The first two premises gain much of their support from the way science is conducted. In the first half of chapter two, Alexandrova carefully argues that we should take the methodological variability we see in the various well-being sciences seriously. From the discussion in chapter one, we learned that different contexts prompt different kinds of evaluations. The kinds of questions a scientist seeks to answer and the resources she has available matter for how she will characterise and measure what she calls well-being. Alexandrova acknowledges that such different approaches could in principle be compatible with, and supported by, the existence of a single unified substantial well-being theory. However, since no theory to fill this function is currently available, she puts her bet on the pluralism running deep.

Alexandrova's reasoning here is persuasive. I find it difficult to disagree with her on this, especially if the toolbox of premise 1 is understood as giving support for constructs, rather than ultimately justifying them in the evaluative sense. A well-being theory could support a construct by pointing to important insights, such as that well-being is perspectival or that people tend to flourish by being virtuous. Different constructs of well-being might be relevant and helpful for differing contexts and types of evaluation. Therefore, even if two incompatible well-being theories cannot both justify (evaluatively) a well-being construct, they could both constitute support for it. Invariantists should accept this.

The Invariantist should instead question premise 3. Why should someone who does not already accept Variantism give credence to the claim that constructs used in the sciences sometimes pick out the constituents of well-being, unless they have scrutinized the constructs in question? In her discussion, Alexandrova acknowledges that constructs often are based on indicators rather than constituents, but she maintains that this need not always be the case. Sometimes, she claims, researchers consider themselves to be investigating the constituents of well-being. She argues that the burden of proof should therefore be on the Invariantist to show that they really are not. The Invariantist, however, could at this point hold their ground and insist that we have no reasons to blankly believe a well-being construct specifies constituents, and especially not to believe that it *both* specifies the constituents *and* gets it right (as premise 4 states), unless we carefully look into the specific construct.

Regardless of how the chips fall on *Theory diversity*, Alexandrova's discussion has at this point led us to an important insight. Scientific constructs of well-being should be justified by their epistemic merits in the contexts they are used. A general substantial well-being theory might be fairly silent on which of the competing constructs is best in a given context. To illustrate, consider *Stated preferences*, a well-being construct sometimes used in welfare economics (see, for instance, Benjamin et al. 2014). Such a construct *could* be supported by a desire satisfaction account of well-being, but not if we believe humans tend to be poor at gauging their own desires (which could be especially plausible if the account invokes idealised desires). At the same time, such a construct *could* also be supported by a hedonistic or objective list account of well-being, if we have reasons to believe humans tend to align their preferences with what benefits them.

Variantists and Invariantists alike should acknowledge that for most practical purposes, we need conceptions that combine the value-making features of substantial well-being theories with our knowledge of human beings and their needs, wants, and desires in different situations of life. It is a shortcoming of the traditional approach to only focus on abstract well-being theories. In chapter three, Alexandrova goes on to propose and exemplify an alternative to this traditional approach by laying out a theory of child well-being. Alexandrova calls this kind of theory a *mid-level theory* of well-being. A central feature of such theories is that well-being is predicated not on individuals, but on kinds. This is an interesting suggestion and I hope the book will mark the beginning of a research program where mid-level theories of well-being will be discussed and developed much further. Well-being scientists often investigate general tendencies that only hold in a specific kind of situation, rather than universal claims. Well-being predicated on kinds might therefore better characterise what they are attempting to measure. At the same time, the focus on mid-level theories should help us see the gap between scientific well-being measures and well-being as it figures in central normative debates concerning the good life, justice and moral value. We should be careful not to forget that it is the well-being of individuals that matters normatively. We care about beings, not about kinds of beings.

In part two of the book, Alexandrova moves on to discuss questions that regard the scientific status of well-being science. In chapter four, she argues that well-being science must be value-laden, but that this is compatible with scientific standards of objectivity. Chapter five makes the case against critics of well-being measurability. Alexandrova discusses an objection she attributes to Hausman (2015), which is that well-being cannot be measured because people are too heterogenous. Even if we can measure some individual well-being components, we cannot know how much different components contribute to different people's well-being. Alexandrova accepts the objection in the case of general all-things-considered evaluations, but she argues that well-being predicated on kinds might still be measurable. Chapter six contains a careful and critical discussion of how well-being constructs are validated in psychometrics.

To wrap up, the overall theme of this book resonates with a development we have seen in other areas of philosophy. Universal aims are being questioned. In political philosophy, proponents of non-ideal theory question the old approach of searching for universal theories of justice. In philosophy of science, the old focus on universal general laws and explanations is being replaced by detailed discussions of less universal mechanisms and local law-like regularities. Alexandrova challenges the old focus on universal and general theories in philosophy of well-being, and argues that they should be replaced by a new focus on mid-level theories.

REFERENCES

Benjamin, Daniel J., Ori Heffetz, Miles S. Kimball, & Nichole Szembrot. 2014. "Beyond Happiness and Satisfaction: Toward Well-Being Indices Based on Stated Preference." *The American Economic Review* 104 (9): 2698–2735.

Cartwright, Nancy, Towfic Shomar & Mauricio Suárez. 1995. "The Tool Box of Science". *Poznan Studies in the Philosophy of the Sciences and the Humanities* 44: 137–149.

Crisp, Roger. 2017. "Well-Being." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Retrieved from <https://plato.stanford.edu/entries/well-being/>.

Hausman, Daniel M. 2015. *Valuing Health: Well-Being, Freedom, and Suffering*. Oxford University Press.

Mats Ingelström is a final year PhD candidate at Stockholm University. He has written and lectured on well-being, measurement and political philosophy. His dissertation project investigates the policy relevance of subjective well-being science.
Contact e-mail: <mats.ingelstrom@philosophy.su.se>

Erasmus Journal for Philosophy and Economics,
Volume 11, Issue 1,
Spring 2018, pp. 107-116.
https://doi.org/10.23941/ejpe.v11i1.358

Review of Yahya M. Madra's *Late Neoclassical Economics. The Restoration of Theoretical Humanism in Contemporary Economic Theory*. New York: Routledge, 218 pp.

RAMZI MABSOUT
American University of Beirut

At the turn of the century, Colander (2000) announced the death of neoclassical and the birth of the *new millennium economics*. The term neoclassical, Colander argues, is a good description of economics around 1900, but it no longer offers an accurate account of contemporary 'modern' economics. According to Colander, the term *neoclassical economics* died when (i) modern economics discarded six of its core attributes, and (ii) replaced them with the *new millennium economics*. This transition has taken neoclassical economics in two distinct directions: one direction is experimental economics and evolutionary game theory; the other direction is complexity theory. Similarly, Davis argues that these new fields "share relatively little in common either with each other or with neoclassical economics" (2006, 1).[1] Colander and Davis thus agree that contemporary mainstream economics is pluralistic.[2] In *Late Neoclassical Economics: The Restoration of Theoretical Humanism in Contemporary Economic Theory*, Yahya Madra (2017) counters these arguments and offers an alternative narrative of the past and present state of neoclassical economics. He also further examines the current state of neoclassical economics in relation to heterodox economics.

According to Madra, neoclassical economics has not been displaced—rather, it is thriving as it exploits recent challenges that re-affirm its core proposition, what he calls *theoretical humanism* (TH). TH is based on two fundamental presuppositions: the first concerns the status of the economic agent, which is a rational, autonomous, self-

[1] Davis (2006) has a slightly longer list with game theory, experimental economics, behavioral economics, evolutionary economics, neuroeconomics, and non-linear complexity theory.

[2] Davis (2006, 10) recognizes that there may be some selection bias (leaving out heterodox approaches from the new pluralism) but concludes "we might say that in recent mainstream economics, though selection bias is no doubt present in some degree, it does not seem sufficient to overcome the pluralist tendencies in the field at the current time".

transparent, and self-conscious individual; the second concerns the search for harmony between individual and social wholes as articulated in the concepts of aggregate rationality and equilibrium. Madra's main thesis is that the new fields that emerged from neoclassical economics (those listed above) also share these two presuppositions. Throughout the book, he identifies these new fields as *late neoclassical economics*, which is defined as the period which followed the post-war neoclassicism, circa the 1970s.

The present state of neoclassical economics is a reaction to the failure of the Arrow-Debreu general equilibrium (A-DGE) axiomatic approach to provide analytic foundations for the discipline. Such foundations were to arise from proofs of existence, uniqueness, and stability of market equilibrium, "from the ground up from individual rational agents" (13).[3] Specifically, late neoclassical economics has three identifying characteristics: (i) it is unified yet heterogeneous; (ii) it is a continuity of A-DGE in its attempt to reconcile individual and social rationality; and (iii) it is a response to the failures of A-DGE.

Madra builds his argument in four parts and ten chapters. Part I offers a summary of the argument and an outline of the Marxist perspective that he draws upon. Part II deals with the problem of structuralism in neoclassical economics (Chapters 3, 4, 5) whereas part III focuses on a selection of late neoclassical topics and how they re-affirm TH (Chapters 6, 7, 8, 9). Part IV concludes (Chapter 10).

In Chapter 1, Madra argues that although contemporary mainstream economics is diverse, it is a partial pluralism that ignores heterodox economics, specifically those approaches that reject TH. Before 1970, neoclassical economics encompassed theories that differed methodologically, ontologically, and politically from each other; however, below the surface, all were committed to TH. Madra argues that the failure of A-DGE—that is, the failure of the second TH presupposition—is misinterpreted by Colander, Davis, and Bowles & Gintis (2000) as a break between post-war neoclassicism and contemporary mainstream economics. He argues that this narrative is misleading because it conflates neoclassical economics with A-DGE and fails to account for other neoclassical traditions including the Marshall-Chicago pragmatic partial equilibrium approach.

Madra then describes how TH underlies neoclassical economics and some heterodox approaches (e.g., the radical political economy approach

[3] All references are to Madra's (2017) book, unless otherwise indicated.

of Bowles 1985). He contends that TH could only be challenged by a rival theoretical orientation, that is to say *structural humanism* (elaborated in Chapters 3 and 10). It is a truism that neoclassicism has changed over the latter half of the 20th century, but these changes, according to Madra, do not amount to a paradigm shift because TH has not actually been abandoned. In fact, neoclassical economics has never stopped cultivating internal heterogeneity, at least since its origins in the different conceptions of equilibrium in Walras and Marshall. What unites all neoclassical economics (early and late) is not a common object of analysis (the market or scarcity) or even a common methodology (mathematical modelling), but TH—it is present across all its economic incarnations.

Any variant of neoclassicism that abandons TH is pushed to the margin of the mainstream. The invisible hand, in the way it aims to reconcile the opportunistic individual with aggregate social harmony via free markets, epitomizes TH. Late neoclassical economics, however, studies market failures and the breakdown of the invisible hand. Rather than being perceived as a failure, late neoclassical economics re-establishes social harmony while acknowledging markets may not be sufficient to achieve the second presupposition of TH. In reaction to troubles in A-DGE, for example, late neoclassical economists have either relaxed axioms or adopted Chicago evolutionary themes. A key attribute of late neoclassical economics is the absence of a "mother structure" as such (95). However, A-DGE remains a point of departure for new fields such as transaction costs, asymmetric information, and game theory.

Chapter 2 starts with a critique of TH, specifically, its atomic anthropocentric element. This anthropocentric vision, inherited from the enlightenment, is not truly secular. A truly secular vision, Madra argues, can be found in the works of Foucault (1983), Althusser (1971), and Resnick and Wolff (1987). The atomic anthropocentric element in neoclassical and late neoclassical economics, better known as *homo economicus*, "functions as the concept of conscious and unified subject that holds together the discipline of economics around the hegemonic reign of the neoclassical tradition" (36). This conception of the agent eliminates the possibility of conceiving the subject as a "site of countless and contradictory influences" (36). While noting that there is general agreement in neoclassical economics on the meaning and definition of rationality, Madra contends TH is itself a point of contention for the various fields that form neoclassical and late neoclassical economics: it is "where the various forms clash with each other in defining the meaning of individual rationality, equilibrium, collective rationality, and efficiency and in determining the correct way to achieve social reconciliation" (38).

Chapter 3 marks the beginning of Part II of the book, subtitled *Neoclassical Economics Under the Shadow of Structuralism*, wherein Madra compares the Marshallian and Walrasian models of equilibrating markets. He argues that the ordinalist turn in neoclassical economics led to the structuralist drifts in A-DGE and the Chicago evolutionary selectionist models of the 1950s and 1960s. Late neoclassical economists preferred the Chicago approach which now forms the foundation of new institutional economics and evolutionary game theory. Madra explains why the Chicago approach was preferred while A-DGE was abandoned: the latter never questioned the efficiency of markets. The Chicago school also naturalized Marshall, generating an all-encompassing social ontology of markets which can be applied to any social phenomenon.

In Part III of the book, Madra focuses on the diverse fields that constitute late neoclassical economics and how they are a continuation of neoclassicism. For example, in Chapter 6, Madra claims that all variants of late neoclassical economics seek to reconcile the failure of the invisible hand, that is, the failure to achieve a harmonious social order. The defining theme here is the study of market imperfection: the continuities and discontinuities between neoclassical and late neoclassical economics are encapsulated in the following three theses: (i) *unity and dispersion*, which claims that late neoclassical economics consists "of a diverse group of economic approaches" that share a theoretical problematic (91); (ii) *continuity*, which claims that late neoclassical economics emerged from neoclassical economics and is not a radical departure from it; and (iii) *response*, which claims that late neoclassical economics is a response to the failures of A-DGE, a response motivated by a desire to rehabilitate TH.

The following three chapters then engage a selection of topics in the new fields of late neoclassical economics, including institutions and information asymmetries (Chapter 7), new ideas about human motivation and limited cognition (Chapter 8), and the proliferation of equilibrium concepts in game theory (Chapter 9).

Chapter 7 opens with the claim that neoclassical economics was criticized for lacking a theory about the internal organization of the firm. The introduction of non-market institutions (e.g., government, firms) in late neoclassical economics stems from three assumptions in neoclassical economics: that agents are unboundedly rational, that contracts are fully specified and enforced, and that a unique, stable equilibrium exists. More specifically, the late neoclassical literature weakens the assumption of fully specified contracts, which is achieved either through the introduction of transaction costs or information failures. While the former is related to the Marshallian-Chicago school, the latter is connected to the A-DGE tradition. Moreover, while both investigate particular conceptions of efficiency, they do not question global efficiency.

Chapter 8 focuses on the human subject, in particular, on motivational diversity and bounded rationality. Late neoclassical economics is a reaction to the cognitive minimalism that dominated up to the 1970s. It responds in two different ways to this minimalism: (i) by supplanting the assumption of opportunism (the "desire to improve one's lot" which leads, according to some late neoclassical economists, to market failures—118) with altruism and reciprocity; and (ii) by acknowledging the limits of human cognition and adopting bounded or procedural rationality. Madra contends that both (i) and (ii) do not constitute a break from neoclassicism, but rather, a rehabilitation of it. The integration of motivational diversity, therefore, does not undermine TH—the individual remains a rational unified, autonomous, and self-conscious. Madra contrasts the literature on motivational diversity to the Chicago pragmatism of Becker (1962) who, without behavioral assumptions, derives downward slopping demand curves from budget constraints. The idea that markets "discipline" re-appears in Vernon Smith (1991) and Charles Plott (1990) where experimental markets are shown to be efficient, notwithstanding the existence of irrational agents. As acknowledged by Madra, Smith and Plott are late neoclassic economists that reject motivational diversity. With respect to limited cognition, Simon's procedural rationality offers a solution to the problems of infinite regress that plague models that assume unbounded rationality or constrained optimization of information à la Stigler. However, Madra argues that since Simon's procedural rationality defines itself by juxtaposition to Cartesian rationalism, it is unable to escape the bounds of TH (in contrast to say Shackle's (1972) *structural uncertainty* where optimization cannot be employed).

Chapter 9 considers the late neoclassical pursuit of equilibrium, the harmonious reconciliation of individual and aggregate rationality. The use of evolutionary arguments in late neoclassical economics is most explicit in game theory where the Nash program faced challenges. To salvage classical game theory, evolutionary game theory deployed Hayek's (1967) concept of spontaneous order. While evolutionary game theory offers causal explanations for motivational diversity, it also reduces the set of plausible Nash equilibria.

Since Austrians reject the second presupposition of TH, Madra inquires whether the introduction of a heterodox concept undermines the TH problematic in late neoclassical economics, or whether, instead, if late neoclassical economics engulfs it in its "gravitational center" to reformulate its TH problematic? Madra favors the latter possibility. Late neoclassical economics can thus "account for the non-coincidence of efficiency and equilibrium without abandoning the normative force of equilibrium even if it is not Pareto optimal" (166). There is only one class of games—disorder games (e.g., rock/scissor/paper)—in which reconciliation is not possible but which are sidelined in late neoclassical concerns. Evolutionary game theory thus reproduces TH to the extent that it models individuals having pre-determined interests that can be reconciled.

Chapter 10 offers an epilogue and contains two sub-sections: the first sub-section explains why the 2008 crash will not generate sufficient criticism of TH in late neoclassical economics; the second offers a non-essentialist Marxist theory of power, one which does not depend upon TH. Madra considers that, given all the investments—intellectual, financial, and institutional—poured into neoclassical economics, its growth and increasing sophistication is to be expected. In its late period, neoclassicism had reached a mature stage from which it is difficult to dislodge, as it strategically employs internal diversity to overcome crises. Neoclassical economics has, in fact, been in crisis since its inception and has shown an ability to absorb criticisms and reformulate its tenets around its TH postulates. Much of the historical critiques of neoclassicism are still effective today. They reappear in the divide between behavioral and experimental economists, whereas the social calculation debate of the first half of the 20th century still divides macroeconomics. The resurgence of Keynesianism after the 2008 crash is further evidence that the old tensions are still present. Neoclassical economics, reborn as late neoclassical economics, is, however, no longer just an intellectual tradition—Madra states that it has turned into "a design for living, a new mode of life, a new governmental rationality, a new model of subjectivity" (178).

Madra's final remarks indicate that his objective is not to challenge the empirical adequacy or logical consistency of neoclassical economics. Instead, he criticizes the claim that there was a paradigm shift between neoclassical and late neoclassical economics, that late neoclassical economics is genuinely pluralistic. Any new synthesis between heterodox and contemporary mainstream economics, at least from his Marxist perspective, is impossible (unless non-TH presuppositions are integrated). His vision distinguishes itself from other heterodox critiques in so far as it is committed

> to produce a knowledge of the social from a perspective that analyzes the different forms of performance, appropriation, and distribution of surplus labor in their irreducibly contradictory and overdetermined relations with each other and with the rest of the social totality. TH is radically opposed to this anti-essentialist Marxian surplus perspective (179).

The Marxist surplus vision does not posit an essentialist subjectivity, nor does it posit micro-foundations for a harmonious social order. Madra's book is rooted in a heterodox tradition which believes that economics is not reducible to, nor able to be reconciled with, TH. *Late Neoclassical Economics* is ultimately concerned that the heterodox critique of the mainstream was undermined and rendered irrelevant by the late neoclassical critique of pre-1970 neoclassicism. It also provides an alternative to appeals made by other heterodox economists commending "a less combative approach than hitherto when trying to win over mainstream economists" (Earl and Peng 2012, 451).

Madra raises many questions that contemporary mainstream and heterodox economists, philosophers of economics, and historians of economic thought will need to ponder and evaluate. While this is beyond the scope of a single book review, I will focus on Madra's identification of TH with neoclassical economics. To make my case, I introduce an illustrative example and then proceed to discuss its implications, which I argue has consequences for Madra's demarcation between contemporary mainstream and heterodox economics. It should be noted that Madra does not contrast 'heterodox' with 'orthodox' but 'heterodox' with 'neoclassical', 'late neoclassical', and 'mainstream economics'.[4]

In the opening pages of the book, it is argued that "heterodox economists are defined by their criticism of mainstream economic theories" (7). It is further argued that some heterodox economists are critical of, at least, one of the tenets of TH (such as Austrian and Sraffaian economists). What defines neoclassical economics is the adoption of the two core presuppositions of TH (a rational, autonomous, self-transparent, and self-conscious individual, and the search for harmony between individual and social wholes articulated in the concepts of aggregate rationality and equilibrium). It is the adoption of these two presuppositions that "distinguish the [neoclassic] tradition from … other, non-mainstream or heterodox traditions in economics" (5).

However, this definition of neoclassicism faces difficulties, as the following example illustrates. The book neglects an emerging late neoclassical field that explicitly abandons the human subject as its atomic agent. This field refers to machines, algorithms, automata, and insects as the ideal neoclassic agents (Binmore 1988; Ross 2005, 2012). Its most vocal spokesperson, Don Ross, rejects "individualism, both methodological and ontological, altogether" (2005, 28). Ross offers arguments that resemble those made by Madra. For instance, Ross asserts that "how neoclassicism (in the version I would call "mature") came to be associated with individualism [is] based on a single philosophical error—taking people as the prototypical agents" (2005, 29). Ross's interpretation, to the extent that it too offers an "anti-anthropocentric view, uniting core insights of neoclassical economics with evolutionary cognitive and behavioral science" (Ross 2005, 19), does not fit Madra's definition of neoclassical economics since it is both anti-anthropocentric and neoclassical. Instead of studying the claims made by Ross and Binmore, Madra focuses on Simon's cognitive economics as well as Davis's (2003) arguments on cyborg economics, and takes Arrow as the exemplar cyborg economist. All of this does little justice to Ross's idea. Instead of confronting Ross's anti-anthropocentric neoclassic synthesis, the cyborg project is dismissed as "a highly contestable proposition" that fails to "liberate" preferences from their anchor in mental entities (55). Madra does not spend much time explaining what is contestable about this proposition. However,

[4] Excluding quotations from other authors, the term orthodox is only used once (179).

since Ross removes any ambiguity about the real aims and ambitions of the cyborg project, the question is whether Ross should be classified as a heterodox or neoclassical economist?

Given his rejection of the first TH presupposition, we could classify Ross's work as heterodox. But this leads to the problematic outcome that Ross's position is not neoclassicist, while he claims it is. So, either Ross is a heterodox economist marketing his view as neoclassical or he is a neoclassical economist that does not satisfy the definition of neoclassicism advanced by Madra.

I will end the review with a thought on pluralism. I am not convinced that contact between heterodox and contemporary mainstream economics ought to be limited to the question of adopting or rejecting TH. A dialectic that, dare I say, fosters multidisciplinary pluralism within economics, that encourages contact between paradigms in terms of the TH problematic but also beyond it, can enrich the discipline.

This was a challenging book to review and I may, in places, not have fully captured the complexity and nuance of the author's view. The period it took to write, over ten years, gave Madra the necessary time to mature his ideas. The breadth of knowledge deployed is impressive and he must be praised for offering a rare detailed analysis of neoclassicism and its subsequent resurgence. Madra ably deploys a critical lens that is both powerful and convincing. I hope it gets the attention it deserves from all quarters of the field.

REFERENCES

Althusser, L. 1971. *Lenin and Philosophy and Other Essays*. Translated by Ben Brewster. New York: Monthly Review Press.

Becker, G. 1962. "Irrational Behavior and Economic Theory." *Journal of Political Economy* 70 (1): 1-13.

Binmore, K. 1988. "Modelling Rational Players II." *Economics and Philosophy* 3 (1): 9-55.

Bowles, S. 1985. "The Production Process in a Competitive Economy: Walrasian, Neo-Hobbesian, and Marxian Models." *American Economic Review* 75 (1): 16-36.

Bowles, S., and H. Gintis. 2000. "Walrasian Economics in Retrospect." *Quarterly Journal of Economics* 115 (4): 1411-1439.

Colander, D. 2000. "The Death of Neoclassical Economics." *Journal of the History of Economic Thought* 22 (2): 127-143.

Davis, J. 2003. *The Theory of the Individual in Economics: Identity and Value.* London & New York: Routledge.

Davis, J. 2006. "The Turn in Economics: Neoclassical Dominance to Mainstream Pluralism." *Journal of Institutional Economics* 2 (1): 1-20.

Earl, P., and T-C Peng. 2012. "Brands of Economics and the Trojan Horse of Pluralism." *Radical Political Economy* 23 (3): 451-467.

Foucault, M. 1983. "The Subject and Power." In *Afterword to Michel Foucault: Beyond Structuralism and Hermeneutics,* edited by H. Drefyfus and P. Rabinow. Chicago: Chicago University Press.

Hayek, F. 1967. *Studies in Philosophy, Politics and Economics*. Chicago: Chicago University Press.

Plott, C. 1990. "Rational Choice in Experimental Markets." In *The Limits of Rationality,* edited by K. Cook and M. Levi. Chicago: Chicago University Press.

Resnick, S., and R. Wolff. 1987. *Knowledge and Class: A Marxian Critique of Political Economy*. Chicago: Chicago University Press.

Ross, D. 2005. *Economic Theory and Cognitive Science*. London: MIT Press.

Ross, D. 2012. "The Economic Agent: Not Human, but Important." In *Handbook of the Philosophy of Economics*, edited by U. Mäki, 691-735. North Holland: Elsevier.

Shackle, G. 1972. *Epistemics & Economics: A Critique of Economic Doctrines*. Piscataway, NJ: Transaction Publishers.

Smith, V. 1991. "Rational Choice: The Contrast between Economics & Psychology." *Journal of Political Economy* 99 (4): 877-897.

Ramzi Mabsout is assistant professor of economics at the American University of Beirut. His research interests include development and welfare economics, policy and ethics in economics, methodology and the history of economic thought, and the philosophy and history of decision theory and reasoning.
Contact e-mail: <rm95@aub.edu.lb>

Erasmus Journal for Philosophy and Economics,
Volume 11, Issue 1,
Spring 2018, pp. 117-124.
https://doi.org/10.23941/ejpe.v11i1.359

Review of Herbert Gintis's *Individuality and Entanglement: The Moral and Material Bases of Social Life.* Princeton: Princeton University Press, 2017, 357 pp.

MICHIRU NAGATSU
University of Helsinki

In his own words, Herbert Gintis's latest book is "an analysis of human nature and a tribute to its wonders" (3).[1] More prosaically, it is a collection of essays, some of which are original and others published elsewhere. Instead of being structured around topics in decision and game theory, like his previous book (2009), this book develops interrelated themes, such as the evolutionary origins of moral sense, its central role in political games, and the socially entangled nature of human rationality and individuality. Some chapters develop Gintis's vision of the unified behavioral sciences by model-building demonstrations; others do so by reflecting on history and methodology.

The demonstrative part of the book models the evolution of human socio-political systems, power relations in markets, altruism, voter turnout, and Walrasian dynamics—drawing on decision theory, game theory, evolutionary theory, and complexity theory. This part offers readers familiar with formal apparatus an excellent overview of the Gintis's recent contributions to the field. The reflective part discusses the nature of rational actor models, provides an intellectual history of sociology and economics, and advocates the unification of the behavioral sciences. This part gives readers interested in history and philosophy of the behavioral sciences an insightful first-hand account by one of the leading figures in the field.

In this review, I focus on Gintis's advocacy of interdisciplinarity, mostly commenting on the reflective part of the book. I also include a meta-review—a review of three reviews of the book, written by other scientists. Regarding the demonstrative part, I strongly recommend the readers to carefully study it themselves, because that is the only way to really appreciate the original insights of the formal work. (And if they are too busy to read all of it, they should at least read Chapters 3 and 9.)

[1] All references are to the book reviewed unless indicated otherwise.

Herbert Gintis is an outstanding veteran figure, who freely travels across boundaries between economics, sociology, anthropology, political science, psychology, and biology. He calls his field "the behavioral sciences", defining them as "the social sciences plus sociobiology (the biological study of the social behavior of living organisms)" (267). I will follow this definition throughout this review. His take on interdisciplinarity is well-summarized in the last chapter (Chapter 12: *The Future of the Behavioral Sciences*), where he observes:

> I have found that when I attack problems concerning human behavior, restricting myself to knowledge from a single academic discipline leaves me partially blind. I find I do much better by combining insights and models from a variety of behavioral disciplines, *letting my research wander about in whatever direction seems fruitful at the moment* (67, my italics).

Gintis sometimes characterizes his approach as transdisciplinary, but, in my view, it is more appropriately captured by what Steve Fuller (2013) calls a 'deviant interdisciplinary' perspective. A deviant interdisciplinarian tries to reverse "epistemically undesirable tendencies inherent in disciplinized inquiry" (Fuller 2013, 1900). In contrast, what Fuller calls a 'normal interdisciplinaran' proceeds by taking this division for granted and assembling disciplines *post hoc*, respecting each participating discipline's expertise. Gintis is clearly not a normal interdisciplinarian—which is a Kuhnian notion—since he denies the maturity of most of the behavioral sciences in the first place. For example, Gintis's insistence that all behavioral scientists should adopt a common core—decision theory and game theory as analytic modelling frameworks—suggests the deviant nature of his approach. This deviant nature will be resisted by many social scientists who do not have expertise in, or appreciation of, formal modelling methods.

The argument Gintis gives for his deviant interdisciplinarity, which is an updated version of the argument in his previous book (2009, Chapter 12) is as follows: the behavioral science disciplines co-exist while holding mutually incompatible models of human nature (or human society, or human behavior). However, since there is only one truth, at most one of these views is correct.[2] Living with this status quo of mutual incompatibility hinders scientific progress. To make progress,

[2] Gintis thinks that, in fact, these views are all false, strictly speaking. The book aims to integrate them by making them correct and mutually consistent with each other.

we must establish a common core analytic framework that serves as (1) a clearinghouse of compatibility between the behavioral sciences, as it were, and (2) a set of shared theoretical templates in which progress can be made in a cumulative fashion.

This argument is based on the assumption that a discipline (which Gintis defines as a common set of questions and communication platforms) is an effective unit of epistemic inquiry to the extent that it roughly coincides with some analytical foundations, such as mathematically formulated evolutionary theory. Since many behavioral sciences lack one, his argument goes, the current disciplinary divisions (in particular, those organized by the departmental system used by many universities) are not epistemically optimal. So, ultimately, Gintis subscribes to the Kuhnian notion of normal science, claiming only that the behavioral sciences have not reached the maturity of a normal science, with the exception of economics. He sounds deviant, but actually is very traditional in this sense.

I have two worries here, one concerning Gintis's optimism about scientists' cognitive capacities, and the other concerning his neglect of underdetermination. My first worry is, more specifically, about his claim that in order to achieve his vision of the unified or integrated behavioral sciences, it is necessary for "the researcher to be fluent in both analytical model building and the thick description of social behavior" (271). Although ideal, there are probably some cognitive constraints on how 'fluent' one researcher can be in both during her career. Interdisciplinary collaboration after disciplinary training is an obvious alternative to the one-person interdisciplinarity, although it poses its own cognitive and institutional challenges. Gintis has collaborated with many researchers and seems optimistic about the prospect of collaboration: "Cross-disciplinary collaboration works well" (267).

My second worry is that Gintis significantly downplays the problem of underdetermination of theory by observation. He notes: "While it is not uncommon for scientists to disagree, there is only one truth in science and standard scientific protocols dictate that disagreements be adjudicated until some resolution is achieved" (268). This is probably the right attitude for a working scientist, but there is often, if not always, room for scientists' epistemic as well as non-epistemic interests to direct them to different, incommensurable or incompatible conclusions (for a case study of incommensurable game theoretic models of social norms, see Paternotte and Grose 2013). This is not about marginal cases in which ideologies create a bogus science—rather, underdetermination is a fundamental aspect of many successful scientific disciplines (Longino 2013; Chang 2012; Mitchell and Dietrich 2006).

Readers of this review might wonder how other behavioral scientists respond to Gintis's deviant interdisciplinarity. In order to get a sense of the responses, I have identified three published reviews of this book in an economics, a psychology and a biology journal. Below, I give a meta-review—a review of book reviews of the book I am reviewing, while at the same time reviewing the book itself.

First, and perhaps most interesting for the readership of this journal, Eyal Winter (2017), who is a professor of economics at the Hebrew University of Jerusalem, gives a big cheer for Gintis's intellectual imperialism:

> I, like Gintis, believe that interdisciplinary research is crucial for real advancements in our understanding of social phenomena. I also believe that it is economics that has to perform most of the courtship in this relationship. Economics is often hailed or blamed for its academic imperialism. [...] For us economists to take the lead on paving the way for interdisciplinary work in the social sciences would be the right thing to do both morally and practically. Morally, because we are the invaders; practically, because economics is primarily about incentives and we need new research incentive schemes within and across disciplines to break disciplinary chauvinism and motivate interdisciplinary research (140).

This comment significantly inflates Gintis's point that other social sciences "are in such serious need of a unifying theoretical framework that a little imperialism from more successful fields should be welcome" (xviii). In turning this little imperialism into a moral duty of the invaders, Winter underestimates how much revision Gintis demands of the standard economic methodology, such as giving up methodological individualism (Chapters 3 and 5); letting go of the selfishness assumption (Chapters 2 and 6); and complementing equilibrium models with dynamic ones in the study of markets while integrating methods and insights from other disciplines along the way (Chapter 11). Some of these changes are surely easier than others, but collectively they may be as demanding as requiring non-economists to adopt decision theory and game theory as their core analytic frameworks.

In stark contrast, Dwight Read (2017), a professor of anthropology at the University of California, Los Angeles, bluntly dismisses the achievements of the book. In addition to the complaint that Gintis does not cite his book, Read makes two criticisms. First, he criticizes the ubiquity of "[t]he attribution of wide-ranging explanatory power to what are simply small pieces of a much larger picture, such as gene-culture co-evolution" (4). This is a somewhat surprising comment given Read's own work, which tries to explain how the increase in short-term memory of our ancestors relates to the evolution of human societies. This seems to me exactly an example of gene-culture co-evolution, which is discussed extensively in Gintis's book (in particular, in Chapters 1, 2, 8 and 10). I suspect what Read really means by "small pieces of a much larger picture" are analytic models, which need to be supplemented by other concepts and field data. But if this is the case, there is no methodological disagreement, because Gintis explicitly notes the importance of "conceptual sophistication in dealing with ethnographic and historical data, as well as a deep feeling for the less formally modeled aspects of social life" (271).

Read's second criticism is that Gintis fails to make a basic anthropological distinction between emic and etic concepts. Emic concepts are those concepts used by the native populations under study to understand their own world (such as gods' will); etic concepts are used by scientists to explain the natives' belief systems and practices (such as the need for social cohesion). Read argues that Gintis's rational model of voter turnout (Chapter 3: *Distributed Effectivity: Political Theory and Rational Choice*) fails because its etic assumption about voters' beliefs and preferences may be different from its emic counterparts. I think that this criticism is misguided, because Gintis does crucially rely on the data about people's reported beliefs and behavior in refuting his rival theories of voter turnout (see Section 3.6). How rational choice models are related to the actors' self-understanding of their own behavior is an important theme in the philosophy of social science, which is also discussed in economics as a use of 'as if' models. Although I cannot discuss this theme in more detail here, I should note that Read's methodological requirement that the etic concepts should coincide with emic ones seems to be too restrictive, especially when one's goal is to explain behavioral patterns, rather than understand how the actors see the world.

Gintis also touches upon the conflation of the emic and the epic (see Sections 3.2 and 6.5). His rational actor model captures the trade-off between three distinct motivations: material (self-regarding), prosocial (other-regarding), and moral (universal). In defending this model, Gintis clearly cautions against assuming that the rational actor model is incompatible with the fact that people have a sense of moral obligations (50). This mistake seems to result from a confusion between emic and epic concepts. Part of the emic sense of moral obligations is that you ought not escape them (at least, not so easily) when the material stakes to do so are high. However, it is a categorical mistake to criticize the epic concept of preferences just because it seems to compromise the emic sense of moral obligations. The question is empirical (whether people in fact trade-off between these motivations) rather than methodological (whether emic and epic notions of preferences must coincide—which they do not have to).

The review by Louise Barrett (2017), a professor of psychology at the University of Lethbridge, who holds a PhD in anthropology, is the most balanced one among the three. This is probably due to her own interdisciplinary background: she works on social cognition of human and non-human primates. While writing that Gintis's imperialistic attitude is "deeply annoying" (937), she admires the fairness with which Gintis synthesizes different camps in the debates over inclusive fitness theory (in Chapter 9), noting in general that "[t]here is room for everyone in Gintis's account" (938). She also rightly notes that Gintis's work on distributed effectivity and cognition (Chapters 3 and 5) is largely in line with the extended (or scaffolded) cognition thesis advanced by the cognitive anthropologist Edwin Hutchins (1995), the philosopher of biology Kim Sterelny (2010), and the philosopher of mind and cognitive science Andy Clark (2008). I would add to this list Wynn Stirling's (2012) game-theoretic approach to the social entanglement of preferences. The convergence of this body of work and Gintis's work on the distributed mind thesis is a recognizable trend in the behavioral sciences that needs more attention from philosophers of science.

Barrett fears, perhaps correctly, that anthropologists and sociologists who are trained in the critical tradition will not be persuaded by Gintis's call for the integrated behavioral sciences because his analysis appears to assume that "our current economic system is somehow inevitable" (938). Of course, this is a sloppy identification of explanation with justification. In fact, there is nothing in Gintis's analysis of economic or political systems (Chapters 4, 7 and 11) that implies the inevitability of capitalism. Chapter 4 (*Power and Trust in Competitive Markets*) analyses the origins of power asymmetry in the market economy, with a modest conclusion that "there is no general theory of when intervention in variable quality markets will enhance economic efficiency" (87). Moreover, Gintis does not sound particularly optimistic about the long-run success of our species:

> Successful cultural changes are often maladaptive (Edgerton 1992), but so far, and in the long run, human culture has been extremely adaptive. Whether this will continue in the face of the proliferation of nuclear weapons, climate change, and reduction in biodiversity remains to be seen (2).

One might criticize this tone of detachment from the urgent emic concerns as an inevitable consequence of adopting a sociobiological perspective on human society, but this association turns out to be wrong: one of Gintis's motivations for integrating the behavioral sciences is to aid in improving socio-economic policy in the areas including "social inequality, poverty, discrimination" (275). So I hope that Barrett's impressionistic reading of Gintis as an economics imperialist and capitalist will not deepen the futile divide between the so-called 'positivist' and 'hermeneutic' camps in the social sciences.

In sum, my meta-review of these three reviews suggests that different disciplines will receive Gintis's call for unified or integrated behavioral sciences in different ways. We have observed the enthusiasm for economics imperialism, the scepticism toward rational choice models, and the pessimism over the book's ability to bridge the ideological divide in the social sciences. Although the sample size is extremely small, I suspect that these reviews simulate some of the typical reactions to Gintis's call for interdisciplinary behavioral sciences. This means that his project will face obstacles in practice, some of which are due to misunderstanding across disciplines, others more substantial. I hope that my critical comments on each reaction will alleviate the first type of obstacles and facilitate the fruitful interdisciplinary discussions that this book deserves.

REFERENCES

Barrett, Louise. 2017. "Uniting the (Social) Sciences?" *BioScience* 67 (10):937–938.

Chang, Hasok. 2012. *Is Water H2O? Evidence, Pluralism and Realism*. Dordrecht: Springer

Clark, Andy. 2008. *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford: Oxford University Press.

Fuller, Steve. 2013. "Deviant Interdisciplinarity as Philosophical Practice: Prolegomena to Deep Intellectual History." *Synthese* 190 (11):1899–1916.

Gintis, Herbert. 2009. *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*. Princeton: Princeton University Press.

Hutchins, Edwin. 1995. *Cognition in the Wild*. Cambridge: MIT Press.

Longino, Helen E. 2013. *Studying Human Behavior: How Scientists Investigate Aggression and Sexuality*. Chicago: The University of Chicago Press.

Mitchell, Sandra D. and Michael R. Dietrich. 2006. "Integration Without Unification: An Argument for Pluralism in the Biological Sciences." *The American Naturalist* 168 (S6): S73–S79.

Paternotte, Cédric and Jonathan Grose. 2013. "Social Norms and Game Theory: Harmony or Discord?" *The British Journal for the Philosophy of Science* 64 (3):551–587.

Read, Dwight. 2017. "An Economist Presents a Rational Behavior Model of Human Behavior from Sociological and Evolutionary Perspectives." *PsycCRITIQUES* 62 (32): Article 6.

Sterelny, Kim. 2010. "Minds: Extended or Scaffolded?" *Phenomenology and the Cognitive Sciences* 9 (4):465–481.

Stirling, Wynn C. 2012. *Theory of Conditional Games*. Cambridge: Cambridge University Press.

Winter, Eyal. 2017. "*Individuality and Entanglement*, by Herbert Gintis, Princeton University Press, 2016." *Journal of Behavioral and Experimental Economics* 68: 140–141.

Michiru Nagatsu is Academy Researcher at the discipline of practical philosophy, University of Helsinki. His main interests include foundations of human sociality and interdisciplinary scientific practices involving economics. Nagatsu advocates more use of empirical approaches in philosophy of science, including historical, ethnographic, bibliometric and experimental methods, and practices some of these methods himself.

Contact e-mail: <michiru.nagatsu@helsinki.fi>

Erasmus Journal for Philosophy and Economics,
Volume 11, Issue 2,
Spring 2018, pp. 125-129.
https://doi.org/10.23941/ejpe.v11i1.360

PHD THESIS SUMMARY: The Measurement of Wellbeing in Economics: Philosophical Explorations

WILLEM VAN DER DEIJL
Doctorate obtained in October 2017
Erasmus University Rotterdam

Wellbeing is a concept that describes how good life is for the person who is living it—a significant type of personal value. By studying human behavior in the context of the scarcity of resources, economics is often concerned with value in general, and personal value in particular. My thesis, *The Measurement of Wellbeing in Economics,* sets out to answer the following question: To what extent is it possible to study wellbeing empirically in economics? In an attempt to answer this question, I analyze the different methodological strands in the economic literature as well as the philosophical debate on the nature of wellbeing.

The question what wellbeing substantively is, is highly controversial and of central concern to a flourishing literature in philosophy, typically divided in three camps: 1) hedonism—identifying wellbeing with the balance of pleasure over pain, 2) desire-satisfactionism—identifying wellbeing with the satisfaction of desires, and 3) objective list theories—listing a plurality of goods that are valuable to us independent of our attitudes towards them. In economics, a variety of approaches have gained a salient position in the empirical literature. In particular, happiness economics, which uses measures of subjective wellbeing; the preference-based approach to wellbeing measurement; and the capability approach, initiated by Amartya Sen.

A significant part of the thesis deals with specific approaches to wellbeing measurement. First, happiness economics has been growing rapidly over the last twenty years, but goes straight against a prominent idea in the foundation of the subfield of economics that deals with wellbeing—welfare economics—namely, that happiness cannot be measured in a way that is interpersonally comparable. While happiness economists generally object that their measures result in reasonable findings, only explainable by the fact that they actually do measure happiness, their approach still is controversial within economics.

Two chapters (3 and 4) focus on methodological issues with measuring wellbeing through happiness. I first analyze one widespread objection against the happiness approach, namely the problem that our aspirations and preferences may adapt to bad circumstances, such that even in prolonged deprivation, people may find happiness, even though their lives are not good for them. While this problem is often seen as an objection against theories of wellbeing that identify wellbeing with happiness, I argue that there is also an alternative interpretation, namely, that in cases of adaptation, people adjust the standards by which they evaluate their own happiness, even though their lives remain equally unhappy. I argue that as an argument against the efficacy of happiness economics, the latter is more plausible and interesting. This implies that even if happiness-conceptions of wellbeing are correct, our ability to evaluate our happiness may be compromised in case we have adapted.

The subsequent chapter (4) also questions the extent to which happiness economics is successful, but this time limits itself to the question whether it is successful as a method to measure happiness itself (rather than wellbeing at large), given our most plausible accounts of happiness. The chapter notes that many happiness economists borrow Bentham's conception of happiness, but do not consider the problems that have been raised in the philosophical literature against this conception. I analyze Mill's criticism of Bentham's conception, and illustrate that taking on board a plausible part of this criticism has significant implications for our ability to rate our own happiness—a crucial assumption for the methodology of happiness economics. Specifically, the criticism is that having qualitatively new experiences changes the way we evaluate, or even understand, our own happiness. This implies that people who have had very different experiences may evaluate the same sense of happiness differently. While I argue that this problem is distinct from the adaptation problem, both problems share that they illustrate a limitation of our ability to evaluate our happiness, such that it can be compared between individuals, or even within an individual over time.

The preference-satisfactionist conception of wellbeing has been central in economic theory but is generally not used to formulate individual measures of individual welfare at large.[1] However, in recent

[1] It is often used to measure the welfare impact of particular changes in people's lives, in, for example, cost-benefit analyses.

years, in response to the developments of happiness economics, some economists have started to develop such general preference-indices of welfare. In chapter 5, I analyze the particular methodological challenges that such approaches are faced with if they aim to be a successful preference-satisfaction measure of welfare. I argue that, while it is in principle possible to successfully arrive at such a measure, a number of central commitments of preference-satisfaction theories of wellbeing are so data-demanding that in practice, satisfying them all is virtually impossible. In particular, unrestrictedness of the preference space and individuality of preferences are such commitments. Moreover, achieving a satisfactory level of measurement, such as ordinal comparability, and interpersonal comparability are features that require much information about individual preference-structures. As a result, measures of wellbeing based on preference-satisfaction are only feasible at the cost of failing to meet some of their central axiological commitments.

A potential alternative to both preference-satisfaction and the happiness approach that I assess is the capability approach. The capability approach is a broad evaluative framework that takes people's actual plurality of doings and beings—their functionings—and our ability to choose them—our capabilities—to be the central evaluative aspect of lives. The measurement of wellbeing is one of the aims of the approach. The capability approach has been formulated as an alternative to both preference-satisfaction approaches and happiness measures, and is committed to the view that our mental states are not always a good source of information about our wellbeing. Moreover, it attempts to incorporate a number of concerns about the plurality of lives in its account, one of which is the fact that certain functionings may be more important to some than to others. In chapter 6 of my thesis, I analyze to what extent these commitments jointly can be realized in the context of wellbeing measurement and argue this is not the case. As a result, the capability approach must either 1) drop its skepticism of measures of wellbeing based on mental-states, 2) deny that different functionings may matter in different degrees to different individuals, or 3) deny that wellbeing is a measurable concept.

Chapter 7 shifts the discussion from specific approaches to the measurement of wellbeing back to the general question how social scientists should develop measures of wellbeing in light of the disagreement about the nature of the concept. It introduces a term, conceptual uncertainty, to describe this difficulty. The chapter reviews some positions about this problem, one of which is to suggest that different scientific practices can select the philosophical position that best suits their field, given the context. Another position suggests that while there is no agreement on the nature of wellbeing, there may be agreement on a large share of goods that either constitute or contribute to wellbeing, which may be used in scientific practice and policy making. I develop an alternative position, which is based on the idea that while it cannot be expected of measures of wellbeing to be uncontroversial, it can be expected that they are not based on conceptions of wellbeing that are incompatible with all major positions on wellbeing in philosophy. I argue that on the basis of this idea, two central widely shared principles can be defended. The first is an affirmation of the personal nature of wellbeing: whatever wellbeing is on a substantive level, what makes our lives good is highly person-relative. A second principle is a denial of the infallibility of our own ability to assess our own wellbeing. While these two principles create a clear tension in the development of wellbeing measures, I suggest that some social scientists are already developing measures that cut across this tension.

In conclusion (chapter 8), the thesis presents a clear challenge for the measurement of wellbeing. While I have argued that wellbeing is person-relative in a substantive sense, I have also argued that our only methods available for assessing people's person-relative wellbeing information, preference and happiness measurement, are fallible in significant ways. Based on the claims defended in the substantive chapters of the thesis, we can formulate a simple argument that denies that it is possible to develop a sound, complete measure of wellbeing across contexts:

1) Regardless of what wellbeing is exactly, either happiness or preference-satisfaction matters intrinsically to wellbeing (defended in chapter 7).
2) Our ability to measure happiness is limited (chapter 3 and 4), and so is our ability to measure preference-satisfaction (chapter 5).
C) There is always a significant part of wellbeing that researchers have limited access to, and hence, wellbeing measures are necessarily incomplete.

At the same time, I suggest that the importance of the concept of wellbeing warrants scientific attention, and that the lack of an ideal measure should not deter scientists from studying the concept. In the end, this thesis is a call for social scientists to take a more pluralistic outlook on the measurement of wellbeing, as no single measure should be seen as a gold standard, and all are fallible.

Willem van der Deijl is currently postdoctoral researcher at the Centre de Recherche en Ethique at the Université de Montreal. Before obtaining his PhD, he studied philosophy and economics at Erasmus University in Rotterdam. Van der Deijl has been an editor of the *Erasmus Journal for Philosophy of Economics* since June 2014. His research focuses on the concept of wellbeing in the social sciences as well as its nature and its significance to questions in ethics and political philosophy. From September onwards, he will be assistant professor in Business Ethics at the Tilburg University.
Contact email: <wjavanderdeijl@gmail.com>