Juha Räikkä
Jukka Varelius (Eds.)

STUDIES
IN APPLIED
PHILOSOPHY,
EPISTEMOLOGY
AND
RATIONAL
ETHICS

SAPERE

Adaptation and Autonomy: Adaptive Preferences in Enhancing and Ending Life

Springer

Studies in Applied Philosophy, Epistemology and Rational Ethics

Volume 10

Series Editor

L. Magnani, Pavia, Italy

Editorial Board

Prof. Atocha Aliseda
Instituto de Investigaciones Filosoficas, Universidad Nacional Autónoma de México (UNAM), Ciudad Universitaria, Coyoacan, 04510, Mexico, D.F.
E-mail: atocha@filosoficas.unam.mx

Prof. Giuseppe Longo
Centre Cavaillès, CNRS - Ecole Normale Supérieure, Paris, France
E-mail: Giuseppe.Longo@ens.fr

Prof. Chris Sinha
Centre for Languages and Literature, P.O. Box 201, 221 00 Lund, Sweden
E-mail: chris.sinha@ling.lu.se

Prof. Paul Thagard
Department of Philosophy, Faculty of Arts, Waterloo University, Waterloo, Ontario, Canada N2L 3G1
E-mail: pthagard@uwaterloo.ca

Prof. John Woods
Department of Philosophy, University of British Columbia, 1866 Main Mall BUCH E370, Vancouver, BC Canada V6T 1Z1
E-mail: john.woods@ubc.ca

For further volumes:
http://www.springer.com/series/10087

Advisory Board

Akinori Abe
Faculty of Letters, Chiba University, Inageku, Chiba, Japan

Hanne Andersen
Department of Physics and Astronomy, Science Studies, Aarhus University, Aarhus, Denmark

Otávio Bueno
Department of Philosophy, University of Miami, Coral Gables, USA

Sanjay Chandrasekharan
Science Education Centre, Tata Institute of Fundamental Research (TIFR), Mumbai, India

Marcelo Dascal
Department of Philosophy, Tel Aviv University, Tel Aviv, Israel

Gordana Dodig Crnkovic
Department of Computer Science and Networks, School of Innovation, Design and Engineering, Mälardalen University, Sweden

Michel Ghins
Institut supérieur de philosophie, Université Catholique de Louvain, Lovain-la-Neuve, Belgium

Marcello Guarini
Department of Philosophy, University of Windsor, Canada

Ricardo Gudwin
Department of Computer Engineering and Industrial Automation, the School of Electrical Engineering and Computer Science, State University of Campinas, Brazil

Albrecht Heeffer
Centre for History of Science, Ghent University, Belgium

Mireille Hildebrandt
Erasmus University, Rotterdam, The Netherlands

Kenneth Einar Himma
Department of Philosophy, Seattle Pacific University, Seattle, USA

Michael Hoffmann
School of Public Policy, Georgia Institute of Technology, Atlanta, USA

Ping Li
Department of Philosophy, Sun Yat-sen University, Guangzhou, P.R. China

Michael Minnameier
Goethe-Universität Frankfurt am Main, Frankfurt, Germany

Margaret Morrison
Department of Philosophy, University of Toronto, Toronto, Canada

Yukio Ohsawa
School of Engineering, The University of Tokyo, Tokyo, Japan

Sami Paavola
Center for Research on Activity, Development, and Learning (CRADLE), Institute of Behavioural Sciences, University of Helsinki, Helsinki, Finland

Woosuk Park
Humanities & Social Sciences, KAIST, Guseong-dong, Yuseong-gu Daejeon, South Korea

Alfredo Pereira
Institute of Biosciences, São Paulo State University (UNESP), São Paulo, Brazil

Luís Moniz Pereira
Centro de Inteligência Artificial (CENTRIA), Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal

Ahti-Veikko Pietarinen
Department of Philosophy, History, Culture and Art Studies, University of Helsinki, Finland

Demetris Portides
Department of Classics and Philosophy, University of Cyprus, Nicosia, Cyprus

Dagmar Provijn
Centre for Logic and Philosophy of Science, Ghent University, Ghent, Belgium

Joao Queiroz
Institute of Arts and Design. Federal University of Juiz de Fora, Brazil

Athanasios Raftopoulos
Department of Psychology, University of Cyprus, Cyprus

Chiaki Sakama
Department of Computer and Communication Sciences, Wakayama University, Japan

Colin Schmidt
Institut d'Informatique Claude Chappe, University of Le Mans, France

Gerhard Schurz
Institute for Philosophy, Heinrich-Heine University, Germany

Nora Schwartz
Department of Humanities, Faculty of Economics, Universidad de Buenos Aires, Argentina

Cameron Shelley
Department of Philosophy, University of Waterloo, Waterloo, Canada

Frederik Stjernfelt
Center for Semiotics, Aarhus University, Aarhus, Denmark

Mauricio Suarez
Department of Logic and Philosophy of Science, Faculty of Philosophy, Complutense University of Madrid, Madrid, Spain

Jeroen van den Hoven
Values and Technology Department, Delft University of Technology, Delft, The Netherlands

Peter-Paul Verbeek
Department of Philosophy, University of Twente, Enschede, The Netherlands

Riccardo Viale
Department of Sociology and Social Research, University of Milan-Bicocca, Milan, Italy and Italian Cultural Institute of New York (Director), USA

Marion Vorms
University Paris 1 Panthéon-Sorbonne, IHPST, Paris, France

Juha Räikkä · Jukka Varelius
Editors

Adaptation and Autonomy: Adaptive Preferences in Enhancing and Ending Life

Springer

Editors
Juha Räikkä
Department of Behavioural Sciences and Philosophy
University of Turku
Turku
Finland

Jukka Varelius
Department of Behavioural Sciences and Philosophy
University of Turku
Turku
Finland

ISSN 2192-6255 ISSN 2192-6263 (electronic)
ISBN 978-3-642-38375-5 ISBN 978-3-642-38376-2 (eBook)
DOI 10.1007/978-3-642-38376-2
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013938366

© Springer-Verlag Berlin Heidelberg 2013
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Contents

Introduction 1
Juha Räikkä, Jukka Varelius

Adaptive Preferences, Autonomy, and Extended Lives 7
Donald W. Bruckner

Adaptation, Autonomy, and Authority 27
Dale Dorsey

"It Won't Be as Bad as You Think:" Autonomy and Adaptation to Disability 49
Jason Hanna

Autonomy and End of Life Decisions: A Paradox 69
Ben Colburn

Gendered Adaptive Preferences, Autonomy, and End of Life Decisions 81
Serene J. Khader

Sour Clinical Trials: Autonomy and Adaptive Preferences in Experimental Medicine 101
James Rocha

Preference Adaptation and Human Enhancement: Reflections on Autonomy and Well-Being 117
Maartje Schermer

Self-Deception, Adaptive Preferences, and Autonomy 137
James Stacey Taylor

Adaptive Preferences and Self-Deception 149
Juha Räikkä

Procedural Exclusion Criteria for Adaptive Preferences 167
Till Grüne-Yanoff, Sven Ove Hansson

Refusing Life-Saving Treatment, Adaptive Preferences, and Autonomy 183
Jukka Varelius

Author Information 199

Chapter 1
Introduction

Juha Räikkä and Jukka Varelius

The value of individual autonomy is a cornerstone of liberal societies. While there are various conceptions of the notion, it is arguable that on any plausible understanding of individual autonomy an autonomous agent needs to take into account the conditions that circumscribe her actions. If she ignores the limitations to which she is subject, she is unlikely to succeed in living autonomously. Yet it has also been suggested that allowing the options open for one to affect one's preferences threatens autonomy (Elster 1985, Colburn 2011, cf. Bruckner 2009).

The idea is often illustrated in terms of the familiar fable of the Fox and the Grapes. In the story, after realizing that it cannot reach some grapes hanging high on the vine, a hungry fox turns away from them and declares them too sour for its tastes. As the fox's preference is suddenly altered in light of what it conceives to be possible for it, the preference appears to be less than fully autonomous. While the fox's case concerns preference change in the face of the closing of the initially most favored option, a relevantly similar kind of adaptation to circumstances can apparently occur when an agent's options remain as they are and when she gains new and appealing options. In these kinds of cases too, an agent's preference could be shaped by her options in a way that could adversely affect her autonomy.

In one form or another, the phenomenon of adaptive preferences has received attention in, for instance, recent discussions on the oppression of women and utilitarian social choice theory (Elster 1985, Khader 2011, Nussbaum 2001, Sen 1984). Yet it has seldom been directly addressed in bioethical debates, and the existing discussion on autonomy would not seem to cover the problems that preference adaptation gives rise to in the bioethical context. Considering that autonomy has a prominent role in contemporary bioethics (Beauchamp and Childress 2009), this shortage is unfortunate. By gathering together previously

Juha Räikkä · Jukka Varelius
Department of Behavioural Sciences and Philosophy
FIN-20014 University of Turku
Finland
e-mail: {jraikka,jukvar}@utu.fi

J. Räikkä and J. Varelius (Eds.): *Adaptation and Autonomy*, SAPERE 10, pp. 1–6.
DOI: 10.1007/978-3-642-38376-2_1 © Springer-Verlag Berlin Heidelberg 2013

unpublished articles on preference adaptation and autonomy, this volume aims to remedy it in two central areas of contemporary bioethics, in connection with human enhancement and in the end-of-life context.

In biosciences, the questions of how to improve our existing mental and physical capacities and how to control the biological processes underlying normal ageing are drawing increasing attention. In the extreme scenarios, the emerging techniques of human enhancement are seen to ultimately free people from the limitations of human nature, thus enabling the advent of immortal physically and mentally supreme post-humans (Bostrom 2003, Harris 2007, Savulescu, Ter Meulen, and Kahane 2011). However, these possibilities are not yet with us and it appears difficult even to conceive what living with capacities significantly different from those we now have would actually be like. Accordingly, the possibility that our preferences for such lives are adapted to what we see as appealing options should not be ignored.

Modern medicine has, of course, already taken significant steps in extending human life expectancy. Yet the quality of the additional years it has made possible has often been found less than satisfactory. Accordingly, and bolstered by the ageing of societies, also bioethical end-of-life questions have recently arisen to new prominence (Battin 2005, Gastmans and De Lepeleire 2010). As people usually consider ending their lives when their most favored option – that of continuing meaningful life without illness or injury – is no longer available for them, the possibility that their end-of-life choices are based on adaptive preferences appears significant.

Thus, if preference adaptation is incompatible with our conducting our lives in the ways we truly want, it also appears to give us grounds to be concerned about our existing preferences towards human enhancement and ending life. Accordingly, the following eleven articles focus on the relationship between adaptive preferences and personal autonomy in ways relevant to assessing how, if at all, preference adaptation affects autonomy in connection with the choices we make as regards enhancing and ending human life.

In Chapter Two, Donald W. Bruckner focuses on preference adaptation and autonomy in the context of human enhancement and in connection with the extension of human life in particular. On the basis of examining central accounts of autonomy and adaptive preferences he argues that, in the sense that should concern us, adaptive preferences and the actions that result from them can be fully autonomous. Bruckner then applies the analysis of autonomy and adaptation that he endorses to the questions of whether the preference to extend one's life would be autonomous and whether preferences one would have within a significantly extended human life would be autonomous. He argues that such preferences, even if adaptive, would be autonomous in the relevant sense, provided that the agent who held them displayed a particular kind of dialogical reflectiveness with regard to the content of those preferences.

In the third chapter, Dale Dorsey asks whether the normative authority of the decisions and actions based on adaptive preferences is affected by their having that particular kind of origin. In light of his investigation into the concept of

normative authority, Dorsey maintains that, while there is good reason to believe that adaptive preferences are non-autonomous, no account of preferential autonomy can preclude all forms of preference adaptation. Dorsey offers two potential solutions to this puzzle. The first is that, insofar as there is no acceptable method by which to distinguish the normatively authoritative preferences from those that lack normative authority, the view that preferences should play a normative role should be rejected. The second is that the question of whether all forms of preference adaptation involve a lack of normative authority should be reconsidered. Dorsey favors the latter solution and suggests that adaptive preferences that are suitably "deep" are normatively authoritative.

In Chapter Four, Jason Hanna considers preference adaptation and autonomy from the viewpoint of disability, focusing on a case of a disabled patient who refuses life-sustaining therapy. After explaining why adaptation is likely to account for some of the preference changes that often follow the onset of disability, Hanna goes on to challenge two arguments purporting to show that mistaken predictions about the effects of disability on quality of life compromise the autonomy of the medical choices of such patients. He suggests that such predictive errors are problematic because of their effect on the decision-maker's well-being, not because they undermine her autonomy. There is nothing distinctively problematic about adaptive preferences, Hanna maintains, and a decision-maker should heed her adaptive preferences when doing so will improve her well-being. Accordingly, Hanna's discussion suggests that legitimate concerns about well-being are sometimes misrepresented as concerns about autonomy.

In the fifth chapter, Ben Colburn assesses the view that respect for personal autonomy implies that people should have control over the time and manner of their deaths. He suggests that, as people usually make end-of-life choices in circumstances they find unacceptable, their preferences are likely to have been affected in ways that preclude voluntary choice. Accordingly, as voluntariness is a central aspect of autonomy, the end-of-life choices people make are, Colburn continues, seldom autonomous. He argues that we should respond to this rather paradoxical result by aiming to change the circumstances in which people make end-of-life choices so that they would always have acceptable alternatives. This, Colburn concludes, would allow people the best possible chance of an autonomous death.

In Chapter Six, Serene J. Khader considers preference adaptation and autonomy in the context of end-of-life decision-making. More precisely, focusing on gendered preferences related to care, Khader asks whether it is sometimes appropriate for medical practitioners to intervene in end-of-life choices that are adapted to unjust social arrangements. She maintains that such choices often qualify as autonomous in what has been deemed a central sense of autonomy, but that we see them as morally problematic despite that. After arguing that changing the pertinent preferences is not an appropriate goal for medical personnel involved in end-of-life decisions, Khader maintains that concern for patient autonomy sometimes warrants intervening in such choices. Intervention is justified, she continues, when the choices are inconsistent with the patient's value system or are

based on misinformation. Khader draws on feminist theories of relational autonomy to show that such interventions by practitioners and family can be autonomy-enhancing rather than paternalistic.

In the seventh chapter, James Rocha discusses preference adaptation and autonomy in connection with clinical trials for terminal diseases, focusing on the case of a patient who could be given a choice between a safe medication and a potentially much more effective but risky experimental medication. He points out that in such a case the patient's choice for the risky option might be based on adaptive preference formation. Given that a physician often does not know whether providing his patient with the choice would lead to preference adaptation, Rocha continues, it is difficult for a physician faced with such a case to give adequate respect for patient autonomy. However, while preference adaptation thus presents problems from the viewpoint of respecting patient autonomy, there are adequate methods for dealing with the difficulties: preference adaptation need not undermine patient autonomy, Rocha concludes.

In Chapter Eight, Maartje Schermer focuses on adaptive preferences, autonomy, and wellbeing in connection with the novel ways of enhancing human capacities that biosciences are expected to produce. She maintains that some preferences we have as regards the human enhancement technology, both for and against the use of the enhancement techniques, are adaptive preferences. As preference adaptation has been deemed detrimental to personal autonomy, Schermer continues, this raises the worry that some of our preferences toward human enhancement may be both non-autonomous and fail to reflect what is truly good for us. However, even if some of our preferences are problematic in those respects, Schermer argues, the difficulties can be overcome by subjecting the preferences to rational reflection. Accordingly, she concludes, the debate on human enhancement can draw valuable lessons from the results achieved in the discussion of preference adaptation.

In the ninth chapter, James Stacey Taylor connects questions of adaptive preferences and autonomy to self-deception. He remarks that in the face of death and serious illness people are prone both to deceive themselves about their possibilities and to adapt their preferences to what they consider possible for them. Should then a concern for patient autonomy, Taylor asks, give way to a greater focus on beneficence in contemporary healthcare? Instead of treating self-deception and preference adaptation separately, Taylor goes on to answer this question by concentrating on adaptive preferences that are based on self-deception. He maintains that persons can be fully autonomous with respect to the decisions that they make on the basis of self-deceptively engendered adaptive preferences. Accordingly, Taylor concludes, even if there were other reasons to criticize it, the emphasis on patient autonomy should not be replaced out of concern for the prevalence of adaptive preferences within healthcare.

In Chapter Ten, Juha Räikkä maintains that, as adaptive preferences and self-deception are often closely connected to each other, the notion of self-deception can help to clarify what happens in some central instances of preference adaptation. Räikkä uses the analysis he provides to throw light on the normative

question of how we should treat adaptive preferences. He maintains that we usually ought to deal with people's beliefs that are based on self-deception and people's adaptive preferences just as we treat their normal beliefs and preferences. However, he continues, the stringency of requirements of autonomy depends upon the context. Accordingly, though considering the more usual cases can provide us with information about life and death choices, Räikkä suggests, in connection with as important choices as those about ending life an acceptable autonomy requirement may preclude adaptive preferences. He concludes by pointing out some practical difficulties that preference adaptation gives rise to in the healthcare context.

In the eleventh chapter, Till Grüne-Yanoff and Sven Ove Hansson aim to qualify the intuition that adaptive preferences should be discounted or ignored when what is good for people is determined by referring to their preferences. The authors maintain that preference theories of good can employ some adaptive preferences and propose a procedural way of demarcating those that should be discounted. Grüne-Yanoff and Hansson use the formal criteria they present to assess a number of exemplary cases and maintain that such criteria rule out those preferences that intuitively should be excluded. They conclude by pointing out the respects in which the framework they propose can be further generalized.

In the last chapter of the volume, Jukka Varelius asks whether a request for euthanasia based on an adaptive preference is autonomous. He aims to answer this question by considering whether such a request meets the requirements imposed by a theory of autonomy that plays a central role in current biomedical ethics. In light of that consideration, Varelius argues that preference adaptation undermines the autonomy of a request for euthanasia. However, he also suggests that in some cases the problems that preference adaptation causes regarding the autonomy of requests for euthanasia can be solved. If a patient is able to critically assess her request for euthanasia and endorses it in light of the assessment, Varelius concludes, the request qualifies as autonomous despite its being originally based on an adaptively formed preference.

These eleven chapters demonstrate that there are different reasoned ways of conceiving the relationship between preference adaptation and personal autonomy in the contexts of human enhancement and end-of-life questions. Especially as the debate on the topics is only at its early stage, this variety can be seen as an advantage rather than as a problem. Hopefully, it will inspire further debate on the themes in the future, discussions that may, but need not, lead to a more unified view on how preference adaptation and autonomy figure in relation to the choices we make about enhancing and ending human life.

We would like to thank Lorenzo Magnani, the Editor-in-Chief of the Springer book series *Studies in Applied Philosophy, Epistemology and Rational Ethics*, and the editorial board of the series for giving us the opportunity to compile this volume, the authors listed above who contributed to it, Carla Bagnoli and Walter Glannon for their help in assessing some of the articles included here, and Marion Lupu for her prompt and careful work with revising our English.

References

Battin, M.P.: Ending Life: Ethics and the Way We Die. Oxford University Press, New York (2005)

Beauchamp, T.L., Childress, J.F.: Principles of Biomedical Ethics. Oxford University Press, New York (2009)

Bostrom, N.: Human Genetic Enhancements: A Transhumanist Perspective. J. Value Inq. 37, 493–506 (2003)

Bruckner, D.W.: In Defense of Adaptive Preferences. Philos Stud. 142, 307–324 (2009)

Colburn, B.: Autonomy and Adaptive Preferences. Utilitas 23, 52–71 (2011)

Elster, J.: Sour Grapes: Studies in the Subversion of Rationality. Cambridge University Press, Cambridge (1985)

Gastmans, C., De Lepeleire, J.: Living to the Bitter End? A Personalist Approach to Euthanasia in Persons with Severe Dementia. Bioethics 24, 78–86 (2010)

Harris, J.: Enhancing Evolution: The Ethical Case for Making Better People. Princeton University Press, Princeton (2007)

Khader, S.J.: Adaptive Preferences and Women's Empowerment. Oxford University Press, Oxford (2011)

Nussbaum, M.: Women and Human Development: The Capabilities Approach. Cambridge University Press, Cambridge (2001)

Savulescu, J., Ter Meulen, R., Kahane, G. (eds.): Enhancing Human Capacities. Wiley-Blackwell, Chichester (2011)

Sen, A.: Resources, Values and Development. Basil Blackwell, Oxford (1984)

Chapter 2
Adaptive Preferences, Autonomy, and Extended Lives

Donald W. Bruckner

2.1 Introduction

An agent's preferences are sometimes formed as a result of her options or, more precisely, her beliefs about her options. It is sometimes claimed that such adaptive preferences are not autonomous, and that, therefore, actions on the basis of such preferences are not autonomous either. Through the examination of some accounts of autonomy and adaptive preferences, I show that the claim that adaptive preferences are not autonomous is highly ambiguous. I argue for a favored interpretation of the claim that adaptive preferences are autonomous, and show that adaptive preferences and the actions that result from them can be fully autonomous in the sense that should concern us. I connect this discussion to debates about life-extending technologies, by applying the analysis I endorse of autonomy to the question whether the preference to extend one's life would be autonomous and to the question whether preferences one would have within a significantly extended human life would be autonomous. I argue that such preferences, even though they would be adaptive, would be perfectly autonomous in the relevant sense, provided that the agent who held them displayed a certain sort of dialogical reflectiveness with regard to the content of those preferences.

Donald W. Bruckner
Department of Philosophy
Penn State University, New Kensington, PA
The United States
e-mail: dwb12@psu.edu

J. Räikkä and J. Varelius (Eds.): *Adaptation and Autonomy*, SAPERE 10, pp. 7–26.
DOI: 10.1007/978-3-642-38376-2_2 © Springer-Verlag Berlin Heidelberg 2013

2.2 Adaptive Preferences

We shall take an adaptive preference to be a preference that is formed as a result of an agent's beliefs about what his feasible options are.[1] The paradigm case is that of the fox. After struggling to reach the grapes he wants to eat, the fox decides he does not want them after all, for they are too sour for his taste. Before he tries to reach them, the fox believes that grapes are within his set of feasible options. That is, he believes he can have the grapes to eat if he tries to get them. After he fails to reach them, he believes that they are not in his feasible set. As a result of this change of belief about his feasible set, he no longer desires the grapes.

In the fox's case, the set of options he believes are feasible undergoes a contraction. But the contraction itself is not what is important. What is important is that it is his belief about the boundary of his feasible set that influences his preference. Mallory the mouse sees birds eating grapes all the time, but has never for a moment thought that she might be able to reach one to eat it. As a result of the belief she has always had that grapes lie outside of her feasible set, she prefers eating the dry and bland wheat grain that falls to the ground. Bartholomew the bunny has never even seen grapes, foraging always for the grass right at his feet, which he believes is the only option he has. When Mallory and Bartholomew are shown the grapes, given a means of getting them, and told of their supreme tastiness, they say that they do not want them and that they are better off with the diets they have had all along. Mallory's and Bartholomew's preferences are adaptive as well, since they are equally formed as a result of their beliefs about the constitution of their feasible sets.[2]

---

[1] This conception of an adaptive preference is a common one in the literature. Bovens, for example, characterizes the adaptation phenomenon as "a shift of desires in response to what one takes to be feasible or morally correct" (1992, 58). Nussbaum characterizes adaptation as "individuals adjust[ing] their desires to the way of life they know" (2001, 78) and as adjustment "in response to a perception of one's circumstances" (2001, 79).

[2] In possible contrast to the characterization I (and others mentioned in the previous note) have given of adaptive preferences, Elster (1983, 113) indicates sympathy with a view of adaptive preferences according to which "reversal could occur simply by going back to the initial feasible set." Since Mallory and Bartholomew do not prefer grapes once they are made feasible, their preferences may not be adaptive on Elster's view. I do not think that reversal upon expansion of the feasible set is a necessary or sufficient condition of the adaptiveness of a preference, or even a frequent correlate. For space and focus, however, I cannot fully explore this issue. Let it suffice to say that among other things, I have in mind the adaptive preferences of some individuals (especially women) in the developing world (and elsewhere) against, for example, education, health care, restrictions on child labor, and meaningful paid work for women outside of unpaid domestic work. These preferences are often formed in response to clear limitations on the agent's feasible set. Development workers have found that these preferences are often quite resistant to change. Rather than leaping at new opportunities that have been made feasible, many of these individuals leap away (see, for example, Fuller (2011), Khader (2011), Nussbaum (2001)). This sort of case indicates that the failure of this reversal-upon-availability condition does not even hint that the preference was not adaptively formed in the first place.

We might worry about the fox, mouse, and bunny. We might worry that, due to the manner in which they were formed, their preferences are not autonomous. Think of a preference as a prescriptive rule for making choices. If one chooses one's own preference, then one decides for oneself what rule of choice to follow. In that case, the preference is autonomous in the literal sense of a *law* that one has prescribed to one*self*. It is one's own preference. If only the fox, mouse, and bunny had had grapes in their feasible sets, one might think, they would have formed and retained preferences for grapes over their actual diets. It therefore seems that the fox, mouse, and bunny did not choose their own preferences, did not, that is, prescribe those rules of choosing to themselves. The rules were imposed upon them from the outside, due to the causal origins of their preferences. Since their preferences have this bastard origin, we might think that their preferences are not truly their own, that they are not fully autonomous.[3]

2.3 Autonomous Preferences

The bastard origin line of thought just rehearsed focuses on the causal mechanisms by which a preference is formed in order to determine whether the preference is autonomous. I shall ultimately argue that this line of thought misdirects our attention from the issues that should concern us when asking whether an agent can act autonomously when acting on an adaptive preference. That is the question of concern in the present inquiry. I shall argue that it is perfectly possible for a preference, such as an adaptive preference, with a bastard origin to be autonomously retained, and that an agent with an autonomously-retained preference acts autonomously when acting on that preference.

The account of autonomy I will use to make these claims draws heavily on a series of papers on autonomy by Andrea Westlund (2003, 2009, 2011).[4] I shall maintain that in order for a preference to count as an agent's own – i.e., to be autonomous – the agent must be willing to engage in a sort of justificatory dialogue about her preference that Westlund describes as dialogically reflective. "[S]he must," as Westlund (2003, 504) puts it, "actually be prepared to answer for [her preference] in justificatory dialogue." The thought is that to count as autonomous, the agent must have a disposition to hold herself accountable to external perspectives critical of her own. This is to say, in part, that not just any

[3] Here and throughout, I assume that *authenticity* or full *ownership* of a preference goes hand-in-hand with *autonomy* of a preference. This assumption is common in the literature on autonomy. It is questioned by Stoljar (2011) (who is partly following Oshana (2007)) and by Bublitz and Merkel (2009).

[4] I was not aware of Westlund's work on autonomy when I developed the defense of adaptive preferences in my (2009). Although I cannot explore the connections, let it suffice to say that in many ways Westlund's account of dialogical reflectiveness is a better or further spelling-out of what I believe I was trying to get at with the notion of reflective endorsement in my (2009).

reason in favor of the object of a preference will do, but that the agent must be willing to engage, and answer to, perspectives other than her own internal perspective.

To see what this willingness to engage in justificatory dialogue amounts to, consider our mouse Mallory who desires to eat grain as a result of grain being the only food available to her. Version 1 of Mallory, Mallory 1, is asked why she has so readily accepted this diet and set her sights on this and no more. Mallory 1 is only able to offer reasons for grain eating from the grain-loving perspective. She is unable to answer for her preference in a way that would answer to the challenges raised by someone who does not share her perspective. So, for example, she tries to justify her preference for grain by appealing to its grainy texture and its good flavor. If this is all she can muster, then she fails to engage in the justificatory enterprise in a dialogically reflective manner, for she fails to be open to the challenges posed from outside of her perspective.

By contrast, consider Mallory 2, who also has the adaptive preference for grain over grapes. Unlike Mallory 1, Mallory 2 gives the question "What's so good about grain?" proper uptake. As Westlund (2003, 512) puts it, "[s]he exercises, both inter- and intra-personally, a capacity for holding herself answerable to external critical perspectives on her choices and conduct." Mallory 1, in effect, only offers reasons for eating grain that assume the validity or legitimacy of the very perspective being challenged. Mallory 2, by contrast, gives other critical perspectives "precisely the kind of normative standing in her deliberations that [Mallory 1] does not" (Westlund 2003, 512). So Mallory 2 might offer a justification by advancing the value of the rodent traditions she is upholding by continuing to eat grain, as a woodworker who uses hand tools might offer a similar tradition-based justification for his preference to use hand tools rather than power tools. Or Mallory 2 might explain how eating grain would violate a commitment she has undertaken, for example, not to take food from other wildlife as long as her own usual sources of food are in good supply. She might simply explain how the desire for grain fits into her first-order orectic economy. For example she might show how foraging for the grain is very easy and efficient and gives her the time she needs to satisfy other of her desires, such as her desire for a college degree which will be achieved only with long hours of study. Although in the end one might not be persuaded by Mallory 2's attempted justification of her preference for grain, her willingness to participate in the justificatory enterprise is what makes her preference her own, autonomous. Again, to lay stress upon an important point, a person does not count as dialogically reflective if she is merely repeating, for example, the cult leader's wacky beliefs or fallacious arguments that are incomprehensible to the uninitiated. She has to be able to engage in dialog with someone who does not share her brainwashed perspective and be able to make it at least dimly plausible to the uninitiated what she sees in the object of her preference.

These examples give us reason to believe that dialogical reflectiveness is necessary and sufficient for a preference to be autonomous, truly the agent's own. It is necessary, as illustrated by Mallory 1, who is not dialogically reflective. Her

preference seems not fully her own because she is unable or unwilling to engage with perspectives external to her own. She is as thoughtful about her preference as a cult member who, once brainwashed, is only able to see the world through a permanently-affixed lens and unable to take an external critical gaze. Those agents are not actively *choosing*, because they literally do not see any options other than the ones they are following. Dialogical reflectiveness seems to be sufficient for a preference to be autonomous, as illustrated by Mallory 2, who is dialogically reflective. Through her willingness to engage with perspectives external to her own, she shows that she is truly making a choice among competing options and deciding for herself what is valuable. We seem to get from her all we need to allow us to conclude that the preference is truly hers.[5] These are just examples. Much of the work in what follows will be with the aim of making it more plausible that dialogical reflectiveness is necessary and sufficient for autonomy.

2.4 Bastard Origins

Surely it will be objected that Mallory 2's preference cannot truly be her own even though she provides a dialogically reflective justification of it once she has it. That is, we might think that even though she can tell us what she sees in the grain over the grapes and answer to external perspectives critical of her own perspective, the preference for the former over the latter is not fully her own. Rather, that preference has been imposed upon her from the outside. The sense in which it has been imposed upon her is that it came about due to her belief about the constitution of her feasible set. She believed that only wheat grain was feasible for her, which it was, so she formed a preference for it. When her circumstances changed and the grain was made available, her belief about her feasible set was updated, but her preference did not change. We might think that the bastard origin of her preference forever affects its legitimacy.

Let us run with this bastard origin thought for a moment. One type of bastard origin is an origin that is necessarily hidden from the agent while it is in operation. Colburn (2011) calls this type of origin a *covert influence*. A covert influence is necessarily hidden from the agent in that if the presence of the influence were revealed to the agent, then the influence would be defeated. For example, subliminal influences are covert in this sense, for a subliminal influence is one that operates beneath the level of consciousness, and its causal efficacy depends on its being hidden from the agent. So, to take Colburn's (2011, 64) example, if during the screening of a film, single-frame images of ice-cream are projected, then the autonomy of any resulting desire for ice-cream is suspect.

[5] Although I take Mallory 1 and Mallory 2 to parallel the examples of Deferential Wife and Anti-Feminist in Westlund (2003) and to support the view that dialogical reflectiveness is both necessary an sufficient for autonomy, Westlund (2009, 28 and footnote 27) only claims that dialogical reflectiveness is necessary for autonomy. She remains officially agnostic on whether it is sufficient.

To be clear, not all adaptive preferences are formed through covert influences and not all preferences formed through covert influences are adaptive.[6] Again, a preference is adaptive in the sense at issue here if it is formed as a result of one's belief about the extent of one's feasible set. But considering what might be claimed to be wrong with a preference formed through the bastard origin of a covert influence will help us see what might be claimed to be wrong with some adaptive preferences.

So what might be wrong with the preference formed through subliminal messaging for some ice-cream? One might simply point to its very origin, and claim that *since* it was formed by a covert influence it must *therefore* not be autonomous or, as Colburn (2011, 61) helpfully puts it, not reflective of the agent's own decision about what is valuable. Similarly, one might claim of an adaptive preference that *since* its formation was a mere artifact of the agent's beliefs about her set of feasible options, the constitution of which is outside of her control, it must *therefore* not be autonomous. The preference was imposed upon her from the outside and is not reflective of her own decision about what is valuable. Instead, the preference came about through an external, possibly alien, causal mechanism. The agent had no part in causing, forming, or deciding upon that preference, so it is not autonomous.

This general line of argument from the bastard origin of a preference to the conclusion that it is not reflective of an agent's own decision about what is valuable is a non-starter. It is a non-starter, because it confuses the causal explanation of a preference with the justification of the preference.[7] It is true that the causal process that resulted in the desire for ice cream was hidden from the agent, so perhaps we should draw the conclusion that the *formation* of the preference was not autonomous. It smacks of the genetic fallacy, however, to say that because it was formed in such-and-such a manner, it *therefore* is not reflective of the agent's values. The question whether the preference is reflective of the agent's current decision about what is valuable is separate from the question of how the preference came about. The question of how it came about is a

[6] This, even though Colburn (2011, 52) may naturally be read as claiming that all adaptive preferences are due to covert influences. At the risk of letting the number of distinctions get out of hand, I will just note that a covert influence is one type of influence that operates through a non-conscious causal process. Just as I claim that not all adaptive preferences are formed through covert influences, I would also claim that not all adaptive preferences are formed through non-conscious causal processes. This is contrary to Elster's (1983, 25) claim that an adaptive preference is formed through "a causal process occurring non-consciously". In the broadest terms, I am claiming that there are at least three distinctions, and they all cut across each other: adaptive/non-adaptive; autonomous/non-autonomous; and conscious/non-conscious. As I go on to argue in the text, for any given preference, however it is classified according to these distinctions, there is the further question whether it is the *retention* or *formation* of the preference that is at issue when we ask whether it is autonomous.

[7] Indeed, this is the upshot of my critique of Colburn (2011) in Bruckner (2011).

causal/descriptive question; the question of whether the preference is autonomous in the sense of being her own here and now is a justificatory/normative question.

John Christman's historical account of autonomy shares the same defect. He claims that in considering the autonomy of a desire it is not "crucial to determine what the agent's evaluation of the desire is at a particular time. What matters is what the agent thinks about the *process* of coming to have the desire, and whether she resists that *process* when (or if) given the chance" (Christman 1991, 10, emphasis added). So Christman's view is that a preference is not autonomous if the agent would have resisted (or did resist) the process by which that preference came to be.

2.5 Observations and Lessons

Genetic or historical accounts such as Colburn's and Christman's are most plausible as accounts of whether a preference was *formed* autonomously. Christman's account in particular claims that we can make this determination by asking whether the agent who holds a particular preference would have resisted, ex ante, the causal process by which the preference was formed. I shall not take a stand on whether such an *ex ante process* account of autonomous preference *formation* is correct. I merely point out that if such an account is a good account, what it is good account *of* is *autonomous preference formation*

Westlund's account, by contrast, is not an account of the autonomy of preference *formation*. It is an account of the autonomy of preference *retention*. Appropriate to the question the account seeks to answer, the account focuses on the agent's *ex post* justification of the *content* of the preference. So it is an *ex post content* account of autonomous preference *retention*.[8]

Which account is best suited to answering whether an agent acts autonomously in acting on a certain preference? Again, I claim, a historical account that claimed that the formation of a preference was not autonomous so *therefore* action on the basis of that preference is not autonomous would commit the genetic fallacy and confuse the question of the explanation of the formation of the preference with the question of the justification of the retention of the preference. We are concerned with whether action on the basis of an adaptively-formed preference is autonomous. My account, inspired by Westlund's, asks under what conditions action on the basis of that preference is autonomous. On my account, we begin with a preference and ask whether action on the basis of that preference is autonomous. That is to say, we take the preference as given; its origin is neither here nor there. If the agent can provide a dialogically reflective justification of the preference, then the preference is autonomously retained and action on its basis is autonomous. If not, then not.

[8] My thinking here was probably influence by Zimmerman's (2003) division between process accounts and content accounts of the distinction between sour grapes preferences and character building.

But it will be objected: does this view not result in unacceptable boot-strapping of a preference from heteronomous to autonomous? Could preferences that result from brainwashing, hypnosis, subliminal messaging, oppression, drugs, and disease not count as autonomous in the same way? This objection, however, gets its force by conflating the two senses of autonomy that I am taking pains to distinguish and that I am claiming are important to prise apart in discussing the autonomy of preferences: autonomous formation versus autonomous retention. There is not a single question to ask about a preference, whether it is autonomous. There are (at least) two: Was it autonomously formed? And is it autonomously retained? It may be that, on some relevant measure of goodness, the best sort of preference is one that is both autonomously formed and autonomously retained, and it should be these to which we give the title "autonomous" full stop. We are concerned here, however, with when action on the basis of some preference is autonomous. I have claimed that such action is autonomous when the preference is autonomously retained, which it is if and only if the agent is prepared to answer for it in justificatory dialogue.

2.6 Christman's More Recent View

Christman (2009) has changed his tune on his earlier account and acknowledged that the *process* by which a preference is formed should not be the sole locus of reflection. He (2009, 146, emphases in original) writes:

> The object of the reflection here, it should be stressed, is not the process of character formation itself, but the trait such formation produces *in light of that development*. That is to say, autonomy requires that one would not feel alienated from the aspect of oneself in question upon reflection *given* the conditions under which that factor came about. This avoids the worries raised about my earlier view where a separate approval of one's developmental history itself was required.

Christman's discussion (2009, 157-159) shows that he conceives this reflection as occurring ex post, after the agent has developed some preference. The reflection is on the preference, *given* the developmental history of the preference. This reflection determines whether a preference is retained, so this new account is an account of autonomous preference retention, an account of which we seek.

Christman's idea, apparently, is that, upon reflection, an agent may repudiate a preference that came about in some way even though he would not repudiate the very same preference if it had come about in some other way. In the former case, the preference is not autonomous. As Christman stresses, the reflection is not on the process by which the preference came about. The reflection is on the preference itself, *given* the way it came about. The idea seems to be that our agent

may repudiate the preference for x over y given that it came about through some chain of causes C_1, but not repudiate the preference for x over y if it had come about through some chain of causes C_2.

Is this not just another way, albeit a less obvious way, of introducing reflection on the causal genesis of the preference, and hence another case of the genetic fallacy? To get around this charge, Christman would have to show that the values for our agent of x and y are legitimately bound up with the way the preference came about. Suppose our agent prefers watching football to watching baseball, but repudiates this preference upon reflection because he realizes that it came about through indoctrination by his abusive father. On one hand, if the agent rejects the preference *merely* due to its causal genesis, then the agent commits the genetic fallacy.

On the other hand, if the agent rejects the preference because watching football turns his stomach now that he realizes it came from his father, then the (dis)value of football is not due to its causal origin. Rather, it is due to its association, which infects the very nature of watching football for our agent. So it is the negative association that legitimately affects the repudiation, not the genesis. This is similar to the case in which someone repudiates a long-standing preference when she learns that the preference was also deeply held by Adolph Hitler, say. In these cases, though, the reflection is on the content of the preference and the developmental history in itself is completely insignificant. If this second hand is how we are to interpret Christman's more recent view, then the view would be very close to the sort of view I want to endorse, which takes the content of the preference apart from its causal history as the appropriate object of reflection.

Unfortunately, Christman (2009, 154-155, 157-158) still emphasizes the history itself of the development of a preference. Again, on his view, a preference for x over y may be legitimately repudiated if it came about through chain of causes C_1 rather than chain of causes C_2, so the developmental history in itself is still clearly central to what he continues to call his historical account. So Christman's more recent view moves in the right direction by backing off of the claim that the developmental history is the sole or primary locus of reflection. However, by continuing to allow that the causal origin of a preference can legitimately affect whether or not the agent repudiates it, and hence whether the preference counts as autonomous, the more recent historical account still commits the genetic fallacy.

2.7 Another Ex Ante Standard for Autonomy

One might be convinced by my argument against historical accounts of autonomy, such as Colburn's and the earlier Christman's, that concentrate on the ex ante view of the process by which preferences are formed. One might, nevertheless, think that the ex ante perspective has some relevance to deciding whether preferences are autonomous, in the following way. Suppose someone proposes to hypnotize

me or otherwise subvert my conscious preference-formation mechanisms so that I come to prefer *x* over *y*. The reason I might object to this process of preference formation is that I deem it unreliable. It is unreliable in the sense that the process risks issuing in preferences that I do not currently want to have – to devote myself and give all of my possessions to a cult leader, for example. The fear of being exposed to hypnosis or other preference-formation mechanisms that subvert my conscious processes is not so much that the subversive process is itself suspicious as that it can result in preferences the content of which I object to. Whether I object to this process of preference formation depends very much on the content of *x* and *y*.

So if someone proposes to hypnotize me so that I come to form a strong desire to steal from the poor to give to the wealthy or so that I develop an obsession with worshipping pine trees, I will not consent. I will not consent for the reason that I object to the content of those preferences. If someone proposes to hypnotize an addict so that he comes to prefer abstinence over indulgence, the addict might not object. Indeed, he might go to some expense in order to be subjected to the hypnosis, because he would welcome a new preference with that content. What this seems to come to is that there is some value in the ex ante perspective on preference formation. But what is of value on this view is not the perspective there on the *process* of preference formation, as earlier Christman and Colburn would have it. What is of value is the perspective there on the *content* of the future preferences.

We have, therefore, a fourth natural candidate for a view of the autonomy of a preference. This is a view that considers the agent's ex ante evaluation of the content of a preference in assessing ex post whether that preference was autonomously formed. Suppose I find myself, ex post, with the preference for *x* over *y* and I want to determine whether that preference was autonomously formed. To make this determination, I ask myself whether I would have objected ex ante to having a preference for *x* over *y*, on the basis of the content of the preference; i.e., on the basis of my ex ante evaluation of *x* and of *y*. If I would have objected, then the preference was not autonomously formed. If I would not have objected, then it was autonomously formed.

Just to associate the ex ante content view with a name, note that Bernard Williams (1973) endorses something similar to this view. More on Williams presently. For now, it will perhaps be useful to summarize in a table the views of autonomous preferences that we have considered. Here, the row labels indicate the perspective from which the reflection is done – ex ante or ex post – and the column labels indicate the object of the reflection – the process by which the preference is formed or the content of the preference.[9]

[9] Putting later Chrisman into the ex post process cell is a bit of an over-simplification but it is accurate to the extent that reflection on the causal history is an essential part of his account.

Table 2.1

	process	content
ex ante	Colburn & Earlier Christman	Williams
ex post	Later Christman	Westlund

2.8 Segue from Williams to Life-Extending Therapies

Bernard Williams (1973), in his well-known paper on the tedium of immortality, relies on a condition for the worthwhileness of an immortal life that is parallel to the ex ante content view of autonomous preference formation. So let us explore Williams' condition for a moment and then use that discussion to motivate the problem of the autonomy of preferences in a significantly extended human life.

In considering the prospect of an immortal life, Williams claims that to be worth living it must be that "the state in which I survive should be one which, to me looking forward, will be adequately related, in the life it presents, to those aims which I now have in wanting to survive at all" (Williams 1973, 91). This condition is key, and Williams insists upon it. Not only are a person's present character and desires the appropriate ones to use in assessing that person's future life, they are the only candidates, Williams says. "For if we … merely wipe out his present character and desires, there is nothing left by which he can judge it at all, at least as something for *him*" (Williams 1973, 93). This present assessment of future options on the basis of the content of those options as judged from one's current perspective is plainly parallel to the ex ante content condition for the autonomy of preference formation. To be clear, Williams is not addressing the autonomy of future desires. But his condition on whether one should want to live in the future as a person with certain sorts of desires is parallel to the ex ante content condition for the autonomy of the formation of preferences. Both conditions evaluate the content of later desires on the basis of one's earlier standards.

There is a further parallel. Just as the question arises whether an immortal life would be worth living, the question arises whether one's preferences would be autonomous in a human life significantly extended through medical technology.

So with this parallel in the problems and this motivation from Williams about how to address them, let us turn to discussing the problem of the autonomy of preferences in a life significantly extended through medical technology.[10]

2.9 The Problem of Preference Autonomy in Long Human Lives

Life-extending therapies are the subjects of increasing philosophical discussion, often under the heading of human enhancement technologies.[11] The issue on which I wish to focus here is whether the preferences humans would have in significantly extended lives would be autonomous, regardless of whether those lives would be extended forever or only by some significant amount. The first observation to make is that many of one's preferences in a significantly extended human life would be adaptive. One reason that many of one's preferences in an extremely long human life would be adaptive is that human ingenuity and invention are constantly creating new options and changing old options into new ones.

Consider only a few lines of work that were not available even 100 years ago: X-ray technician, airport ground crew member, computer programmer, truck driver. Consider also pastime pursuits that engage people today but that were also not available until relatively recently: download and play music from the internet, watch a movie at a theatre or through Netflix at home, ride a bicycle, play a game of Trivial Pursuit, use a computer to socialize or engage in scholarly collaboration. These lines of work and other pursuits have been created due to human ingenuity and invention.

Not only have new pursuits been created, but old pursuits have been changed into new pursuits. For just one example, consider the nature of the practice of medicine in the 17th century in comparison to the practice of medicine in the 21st century. The nature of the practice of medicine has changed significantly. Advances in science have led to a much greater understanding of human diseases. Such scientific advances have also made diagnosing illnesses a completely different enterprise, for instance, through blood tests and brain scans. Doctors are now able to treat formerly deadly diseases with a range of effective treatments developed only relatively recently, including antibiotics and chemotherapy. As well, physicians now can specialize to a degree that was impossible in the 17th century. The upshot is that the nature of medical practice has drastically changed in 400 years, due to human invention and ingenuity. The change in the nature of the practice of medicine is an example – and but one of many examples that could

[10] In Bruckner (2012) I consider Williams' argument for the claim that an immortal life would be boring, and offer several arguments against his claim, one of which overlaps with the discussion in the next section.

[11] Some entries into the literature include Allhoff, Lin, and Steinberg (2011), Bostrom (2008), Temkin (2008), Harris (2007, esp. Ch. 4).

be given – of how human ingenuity and invention have changed old pursuits into new pursuits.

We can expect, therefore, that in very long human lives, long-living humans will develop preferences for new pursuits that have either evolved out of old pursuits or that have been freshly invented. Such preferences are adaptive in the sense defined at the outset, for an adaptive preference is one that is formed as a result of an agent's beliefs about what her feasible options are. When an agent is cognizant of the expansion of her set of feasible options, then the agent's beliefs about what are her feasible options are updated as well. And if the agent forms a preference for an option of which she is newly cognizant, then this preference is adaptive. The examples I have just quickly surveyed suggest that a human life significantly extended through medical technology would see the introduction of many new options and that we would form preferences in favor of those options. Thus, we can raise questions about the legitimacy, in particular, the autonomy of such adaptive preferences in a significantly extended human life.

It should be clear enough what Williams would say about the prospect of living an eternal life, or just a very long life, with such preferences. A very long life will see the introduction of many options into one's feasible set and the resulting preferences for those options may be inconsistent with one's ex ante preferences, or at least not clearly related to one's ex ante preferences. Such a life fails Williams' central condition for being attractive from one's current perspective, namely, that it must be "adequately related, in the life it presents, to those aims which [one] now [has] in wanting to survive at all" (Williams 1973, 91). With such radical change in the options open to me and the introduction of new preferences for some of those options, the concerns of many of my ex post selves will be quite disparate from those of my ex ante self. So on the basis of my ex ante self's evaluation of the content of my ex post self's preferences, I should reject a significantly extended human life as not worth living, according to Williams.

There is another evaluative dimension on which one might think that an extended human life should be rejected. One might think that the adaptive preferences in a very long human life would not be formed autonomously. And one would think this if one held the ex ante content view of autonomous preference formation. For according to that view, the content of one's ex post preferences should be evaluated according to one's ex ante standards. If that ex post content would be negatively evaluated from one's ex ante perspective, then those preferences are not autonomously formed. If we do not wish to lead a life with non-autonomously-formed preferences, then we should reject a significantly extended life.

2.10 The Irrelevance of the Ex Ante Content Condition

The arguments just given – on behalf of Williams' claim that a very long human life would not be worth living and on behalf of the claim that our preferences in such lives would not be autonomous – are incomplete. In this section, I will

explain why those arguments are incomplete. I will then offer what appears to be the most plausible way of filling out those arguments. Finally, I will explain how that way of filling out those arguments shows what is wrong with Williams' condition and the ex ante content condition.

The argument given on behalf of Williams is incomplete because it merely asserts without justification that our ex post preferences in a very long or immortal life would be at odds with our ex ante preferences and so the concerns of one's ex ante and ex post selves would be quite disparate. Similarly, the argument claiming that our later preferences in very long human lives would not be autonomous needs an argument for the premise that our ex ante and ex post preferences in such a life would be inconsistent. What justification can be given for this claim that both of these arguments need? Of course, we cannot make any direct observations from experience, such as cases of people who lived 140 years and who developed preferences at age 120 to which their 50-year-old selves would have objected. The best we can do is to make observations about the patterns of our preferences in our relatively short lives and make inferences about the patterns that would emerge in significantly longer lives.

I suggest that there is indeed copious evidence of ex ante and ex post preference inconsistency in our short lives and that this evidence provides strong support for the claim that there would be this sort of inconsistency in very long or immortal lives. As a 17-year-old, I preferred commuting to my job on a motorcycle. I now have a strong preference in favor of commuting by automobile. As a younger man I had preferences for a certain kind of work and family life that have changed substantially. As well, I currently have preferences, for instance to produce decent philosophy, that I realize will, later in life, likely fade or give way altogether to other things to which I now assign much lower priority or reject altogether, for instance to spend much of my time in leisure activities that now horrify me, such as reading low-quality novels while sipping cocktails on the deck. To be sure, these are banal autobiographical examples, but I do not think that my case is unique. This sort of example serves to call our attention to the fairly ubiquitous phenomenon of changing tastes and preferences over the course of our relatively short lives. The important point is that it is commonplace, in our brief lives, to hold preferences ex post the content of which our ex ante selves would have rejected.

If this is so for our short lives, then *a fortiori* it would be for lives significantly extended by technology. I say *a fortiori* because none of the examples of preference change that I just gave rely on human invention and ingenuity and the attendant expansion of our feasible sets, but we observed that in long human lives we could expect our feasible sets to change considerably. Consider someone raised in the era of music recordings on vinyl discs. Suppose she once held a preference to listen to music on that medium over all others, but has now come to prefer to listen to music by using an MP3 player. Similarly, many other things will be invented in the next 30 years which will lead us to develop preferences that are inconsistent with our current preferences. So in very long lives not only would we have the sort of preference changes that occur in our short lives and that do not

depend on changes in our feasible sets, we would have additional preference changes as a result of pronounced expansions of our feasible sets. So we will have all of the ex ante/ex post inconsistency that we observe in our short lives and quite possibly more, due to the longer duration of lives and the possibility of developing preferences for formerly infeasible options, preferences that conflict with those we hold ex ante.

That, then, is the further spelling out of the argument on Williams' behalf that very long human lives would not be worth living and the argument on behalf of the proponent of the ex ante content view of autonomy that many of our preferences in very long lives would not be autonomous. I now wish to claim that this further spelling out shows Williams' ex ante approval condition on the worthwhileness of eternal lives to be implausible and the ex ante content view of autonomy to be irrelevant to our concerns about autonomy. The further spelling out moved from the sort of preference changes we experience in our short lives to the very same sort of preference changes we would have in much longer lives, plus some other preference changes. The important thing to note about the sort of preference change we experience in our short lives is that we do not think that having many ex ante/ex post preference conflicts makes our short lives not worth living. So the obvious question for Williams is why we would want to hold very long lives to a standard that is not even satisfied in our short lives.[12] We should not be concerned that the standard would be violated in long lives, because we are not and should not be concerned that the standard is violated in our short lives.

The argument concerning the ex ante content condition for autonomous preferences in long lives is similar. Consider again the audiophile who now prefers listening to music on MP3s rather than vinyl discs. That preference is adaptive, for it came about through the expansion of her feasible set. That preference is inconsistent with the preference she held decades ago in favor of listening to music on vinyl over all other media. Yet the inconsistency between her present preference and her past preference does not lead us to think that she suffers from an autonomy deficit in light of her present preference to listen to music on MP3s. To be sure, I am not actually rejecting the ex ante content condition for autonomous preference formation. The force of my argument is that even if we want to say that such preferences are not autonomously formed because they fail the ex ante content condition for autonomous preference formation, this consideration is not relevant to the question whether we act autonomously when acting on such preferences. We should not think, that is, that failure of the ex ante content condition means that the ex post preferences are non-autonomously retained. So in asking whether we act autonomously on such ex post preferences, it is irrelevant whether they are formed autonomously according to the ex ante content standard. And the basic reason I have offered is that our ex ante preferences and ex post preferences can be quite inconsistent in our relatively short lives, and we do not judge our actions based on our ex post preferences less autonomous as a result.

[12] Fischer (1994, 267-68) and Rosenberg (2006, 232-235) also point out this dual standard.

I have provided parallel treatments of Williams' condition on the worth of an immortal life and the ex ante content view of autonomous preference formation because they stand and fall together. Both conditions consider whether one endorses, ex ante, the content of one's ex post preferences. If not, then Williams' condition says that an immortal life with such inconsistent preferences is not worth living and the ex ante content view says that the ex post preferences were not formed autonomously. I have argued that Williams' standard is not even satisfied in perfectly worthwhile mortal lives, so it should be rejected for very long lives and immortal lives. While I do not reject the ex ante content view of autonomous preference formation, I have stressed that it asks the wrong question if our concern is knowing whether the agent acts autonomously in acting on the preference he has at the time he acts on that preference.

This is why Westlund's ex post content view of autonomous preference retention is the appropriate view to apply. It asks whether the agent is open to justificatory dialogue about her preference now, when she is acting. That her ex ante self does not approve of the content of the preference on which her ex post self acts might matter to whether the preference was formed autonomously. This should not matter, however, to the autonomy of the retention of the preference and the autonomy of the action. All that disapproval from the ex ante perspective implies is that acting on her ex post self's preference would not be autonomous for the ex ante agent. It is not the ex ante agent's preference, so of course acting on it would not be autonomous for the ex ante agent. Ex ante disapproval does not imply, however, that the ex post self's action on the basis of that preference is not autonomous.

2.11 Another Questionable Preference

Perhaps I have established that in a very long human life made possible through life-extending technologies, the preferences one would have, adaptive though they may be, could be perfectly autonomously retained. Still, one might wonder about the status of the preference to extend one's life significantly in the first place. There is currently no enhancement available that would guarantee, or even make highly probable, an extension of a human life by, say, 50 years. Suppose that a pill were to become available that would do just that. We can expect that many people would form a preference to take the pill. The preference to take the pill would be an adaptive preference, arising, as it would, as a result of the introduction of a new option into one's feasible set.

Would such a preference be autonomous?[13] We know by now that this is a trick question, for the question is ambiguous between asking whether the preference is

[13] There is some discussion in the human enhancement literature about the autonomy of preferences for or against enhancement therapies in general, as well as the autonomy of the preferences one has and the actions one performs after enhancement. For one view of the issues involved, see Bublitz and Merkel (2009). See also Harris (2007).

autonomously formed and whether the preference is autonomously retained. I have not plumped for any account of the autonomy of preference formation. I have only said that two possible candidates for determining whether a preference is autonomously formed are the ex ante accounts that consider the agent's reflection on either the content of the potential preference or the process by which the preference is to be brought about. I am officially agnostic on whether or not the adaptive preference to take a pill to extend one's life by 50 years is autonomously formed.

The sense of the question I am prepared to answer, then, is whether the adaptive preference to take the pill is autonomously retained, which is equivalent to the question whether the agent acts autonomously in acting on the adaptive preference to take the pill. I see no reason why the same account I have advocated for other cases would not apply to the case at hand. If, from the ex post perspective of having the preference, the agent evaluates the content of the preference favorably and is dialogically reflective about this evaluation, then the preference is autonomously retained and the agent acts autonomously in acting on it.

Some readers will have a nagging feeling and object, on two possible grounds. First, the preference for taking the pill may have come about due to manipulative advertising, brainwashing, or other suspect causal mechanisms that make it clear that there was nothing agentive about the acquisition of the preference. In such a case, even if the agent endorses the content of the preference upon reflection, the reflection itself may be handicapped by, or be an artifact of, the state into which the agent has been manipulated, brainwashed, or otherwise had her agency removed through the formation of the preference. So we should not want to say the preference is autonomous even though the content is endorsed from the ex post perspective. Second, if the preference to take the life-extension pill is not autonomous, then the autonomy of other preferences that come about as a result of this preference should be called into question as well. The idea is that if the preference for taking the pill is non-autonomous, the other preferences that are formed as a result of the options that open up as a result of taking the pill and gaining years of life – for instance, the preference to learn a couple more languages or build a house from scratch by hand – inherit the non-autonomy of the preference to take the pill.

In reply, I want to say that the first part of this objection is not an objection to the view that I have advocated in this paper and that the second part of this objection simply engages the genetic fallacy again. Indeed, I agree with the first part of the objection that if an agent is manipulated or brainwashed into having a preference, then merely endorsing the content of the preference from that brainwashed perspective is insufficient for the preference to be autonomous. Rather, as I pointed out while discussing Westlund's view, the agent must be dialogically reflective. She must be genuinely open to considering objections to the content of her preference from perspectives other than the perspective into

which she has been brainwashed. She must be able to offer reasons that have some hope of engaging someone with a critical perspective other than her own. Merely to offer reasons that only make a brainwashed desire comprehensible to others who have been similarly brainwashed is not to be dialogically reflective. Rather, the agent's attraction to the object of the preference must be able to be made comprehensible to others who do not share her attraction. It is plausible to suppose that someone with a brainwashed preference to take the 50-year life extension pill could make the attraction comprehensible to someone who does not desire to extend his own life through this means. So it seems plausible that the desire to take such a pill could be autonomous for many people, regardless of how they came to have the preference.

With this, the second part of the objection is put to rest, because it depends on the assumption that the preference to take the pill is non-autonomous. While the preference to take the pill might be non-autonomously formed (through brainwashing, for example), I have insisted that it does not follow that it is non-autonomously retained and that action on its basis – and other preferences formed on its basis – are therefore non-autonomous. Further, even if the preference to take the pill were non-autonomously retained, it does not follow that preferences formed as a result of acting on that preference are themselves non-autonomously retained. Suppose I take the 50-year life extension pill at age 70. At age 110, *any* preference I express has come about as a result of acting on the non-autonomous preference to take the pill, for if I had not taken the pill I would have no preferences at all, since I would not be among the living. So on the view of the objector according to which the non-autonomy of the earlier preference is inherited by later preferences, the preference I express on my 110th birthday for chocolate cake with vanilla icing rather than grasshopper cake with chicken feathers on top is non-autonomous, even if I am dialogically reflective about this preference, as we might expect me to be. Along with the grasshopper cake, the view that *none* of my preferences as a 110-year old are autonomously retained, and so none of my actions are autonomous, seems too hard to swallow.

2.12 Conclusion

I have argued that acting on an adaptive preference is autonomous provided that the agent is willing to give a dialogically reflective justification of the preference. This ex post perspective on the content of a preference is the appropriate one to take if our concerns lie in assessing the autonomy of an agent's action on the basis of that preference. I have argued that to ask whether an adaptive preference was formed autonomously is to ask a separate question from the one of concern in asking whether the agent autonomously retains her preference and thereby acts autonomously, which we can answer through Westlund's approach. I connected this discussion of adaptive preferences and autonomy to questions about the status of preferences in significantly extended human lives. I argued that the presence in

an extended life of adaptive preferences should not worry us and lead us to claim that such a life would not be worth living or that acting on those preferences would not be autonomous.[14]

References

Allhoff, F., Lin, P., Steinberg, J.: Ethics of Human Enhancement: An Executive Summary. Sci. Eng. Ethics 17, 201–212 (2011)

Bostrom, N.: Why I Want to be a Posthuman when I Grow Up. In: Gordijn, B., Chadwick, R. (eds.) Medical Enhancement and Posthumanity, pp. 107–137. Springer, New York (2008)

Bovens, L.: Sour Grapes and Character Planning. J. Philos 89, 57–78 (1992)

Bruckner, D.W.: Against the Tedium of Immortality. Int. J. Philos Stud. 20, 623–644 (2012)

Bruckner, D.W.: Colburn on Covert Influences. Utilitas 23, 451–457 (2011)

Bruckner, D.W.: In Defense of Adaptive Preferences. Philos Stud. 142, 307–324 (2009)

Bublitz, J.C., Merkel, R.: Autonomy and Authenticity of Enhanced Personality Traits. Bioethics 23, 360–374 (2009)

Christman, J.: Autonomy and Personal History. Can J. Philos 21, 1–24 (1991)

Christman, J.: The Politics of Persons. Cambridge University Press, Cambridge (2009)

Colburn, B.: Autonomy and Adaptive Preferences. Utilitas 23, 52–71 (2011)

Elster, J.: Sour Grapes: Studies in the Subversion of Rationality. Cambridge University Press, New York (1983)

Fischer, J.M.: Why Immortality is Not So Bad. Int. J. Philos Stud. 2, 257–270 (1994)

Fuller, L.: Knowing Their Own Good: Preferences and Liberty in Global Ethics. In: Brooks, T. (ed.) New Waves in Ethics, pp. 210–230. Palgrave Macmillan, Basingstroke (2011)

Harris, J.: Enhancing Evolution: The Ethical Case for Making Better People. Princeton University Press, Princeton (2007)

Khader, S.: Adaptive Preferences and Women's Empowerment. Oxford University Press, Oxford (2011)

Nussbaum, M.: Symposium on Amartya Sen's Philosophy: 5 Adaptive Preferences and Women's Options. Econ. Philos 17, 67–88 (2001)

Oshana, M.: Autonomy and the Question of Authenticity. Soc. Theory Pract. 33, 411–429 (2007)

Rosenberg, J.F.: Reassessing Immortality: The Makropulos Case Revisited. In: McDaniel, K., Raibley, J.R., Feldman, R., Zimmerman, M.J. (eds.) The Good, the Right, Life and Death: Essays in Honor of Fred Feldman, Ashgate, Burlington VT, pp. 227–240 (2006)

Stoljar, N.: Autonomy or Authenticity? Commentary on Andrea Westlund's 'Rethinking Relational Autonomy' and Catriona Mackenzie and Jacqui Poltera's 'Narrative Integration, Fragmented Selves and Autonomy. Symposium on Gender, Race and Philosophy (2011), http://web.mit.edu/sgrp

[14] An earlier version of this paper was presented to the Philosophy and Religious Studies Department at the University of Mississippi. I am indebted to the audience for helpful feedback. Thanks also to Jukka Varelius and Juha Räikkä for very helpful feedback on an earlier version, which substantially improved the final product.

Temkin, L.: Is Living Longer Living Better? J. Appl. Philos 25, 193–210 (2008)
Westlund, A.C.: Selflessness and Responsibility for Self: Is Deference Compatible with Autonomy? Philos Rev. 112, 483–523 (2003)
Westlund, A.C.: Rethinking Relational Autonomy. Hypatia 24, 26–49 (2009)
Westlund, A.C.: Reply to Benson, Christman, Rocha, and Stoljar. In: Symposium on Gender, Race and Philosophy (2011), http://web.mit.edu/sgrp
Williams, B.: The Makropulos Case: Reflections on The Tedium of Immortality. In: Williams, B. (ed.) Problems of the Self, pp. 82–100. Cambridge University Press, Cambridge (1973)
Zimmerman, D.: Sour Grapes, Self-Abnegation, and Character Planning. Monist 86, 220–241 (2003)

Chapter 3
Adaptation, Autonomy, and Authority

Dale Dorsey

3.1 Introduction

Adaptation is a fact of life. If I walk down the street and fall down a manhole, I'm likely to keep a lookout for open manholes in the future. In this case, I *adapt* to my circumstances; I take a more cautious approach than I might otherwise have. This is true of all forms of life, and plausibly helps to explain the fact that there is, in fact, life at all.

Adaptation of *preferences* is not much different. If I have a preference for something that I can't get, I'm likely to adapt: I will either stop preferring that thing, or perhaps (depending on the form my adaptation takes) come to reverse my preferences. If I desire to go to Harvard and not Columbia, but I get into Columbia and not Harvard, I may come to revise my preferences toward Columbia, and away from Harvard. There could be many reasons for this adaptation, but one chief reason is that to maintain the former preference is painful: it is the frustration of something I want. Perhaps not as painful as falling down a manhole, but a pain no less worthy of psychological adaptation to avoid.

This is a simple fact of human life, unremarkable except for the fact that it causes widespread–and widely noted–problems in moral theory. Preferences are supposed to represent, broadly speaking, a person's *good*; they are the object of our beneficent concern; and, perhaps somewhat more controversially, the proper index by which to measure social choice. But if preferences can be adaptive in this way, each of these roles appears to disintegrate. It would be wrong, for instance, to say that I'm doing better to the extent that I'm going to Columbia rather than

Dale Dorsey
Department of Philosophy
University of Kansas
Lawrence, KS
The United States
e-mail: ddorsey@ku.edu

J. Räikkä and J. Varelius (Eds.): *Adaptation and Autonomy*, SAPERE 10, pp. 27–47.
DOI: 10.1007/978-3-642-38376-2_3 © Springer-Verlag Berlin Heidelberg 2013

Harvard, that the person with a beneficent interest in my welfare would promote a state of affairs in which I attend Columbia *rather than* Harvard, or that a proper social theory should design institutions or policies with an eye toward this preference. The problem gets worse, of course, when individuals adapt to conditions that social policy should generally try to avoid, such as poverty or oppression.

The existence of adaptive preferences might cause one to believe that preferences should play no role in normative domains to which they may be applied.[1] But this inference is too quick. After all, in cases in which a person's preferences are *not* adaptive, it would seem implausible to hold that proper accounts of autonomy, welfare, beneficence, etc., should make no reference to an individual's preferences. But if this is right, a crucial question for ethics and social theory is: is there a principled method by which to distinguish normatively authoritative preferences, and to expunge those that are problematically adaptive?

In this essay, I investigate the connection between adaptation and normative authority. However, my investigation into the concept of normative authority reveals something of a puzzle. Though there is good reason to believe that adaptive preferences are, broadly speaking, non-*autonomous*, I conclude that there is no account of preferential autonomy that can plausibly eliminate all forms of preference adaptation. Hence, I offer two potential solutions to this puzzle: first, that we should simply reject the claim that preferences should play a normative role (insofar as there is no acceptable method by which to distinguish the normatively authoritative ones from those that lack normative authority), or, second, that we should reconsider whether all forms of preferential adaptation signal a lack of normative authority. Briefly, I argue in favor of the latter, and hopefully more optimistic, conclusion.

3.2 Varieties of (the Lack of) Normative Authority

It seems right to say that the fact that a particular preference for φ rather than ψ is adaptive is a reason to look askance at its normative authority. We generally refuse to treat it as a guide to a person's welfare, refuse to take it seriously in planning our beneficent actions, and hold that social choices should not be indexed to the fulfillment of such a preference, etc. (Whether this is actually true will turn out to be a relatively complicated matter to which I will return at the end of the paper.)

But I should distinguish the problem of normative authority that results from problems of adaptation from the problem of normative authority that results from other sorts of preference failures. Adaptation is not the only feature of a preference that generates a lack of normative authority. Preferences can lack normative

[1] See, for instance, Nussbaum (2000, Ch. 2) for a critique of preferences rather than *capabilities* as the proper index of social choice. I criticize Nussbaum's position in Dorsey (2012).

authority because they are *sadistic* (see, e.g., Rawls 1971, 30-31). If I prefer to harm you, for instance, it might be that this preference displays a normative failure. In addition, preferences can lack normative authority because they are *trivial* or, indeed, directed toward *the objectively worse* (see, e.g., Brink 2008 and Kraut 1994). Not all agree that such preferences lack authority. But we should admit that adaptation *per se* should be distinguished from failures of authority for such reasons. I can maintain a sadistic preference non-adaptively. I can adaptively prefer something that is sufficiently objectively good, etc.

For the purposes of this paper, I want to focus on problems of normative authority that stem specifically from adaptation. This is not to say the remaining problems are not particularly serious. But they are, I think, important to keep distinct.

3.3 Adaptation and Autonomy

What distinguishes the *per se* failure of normative authority associated with adaptation? What, more to the point, *explains* why adaptive preferences are not authoritative (in the way that, say, a bad will explains why sadistic preferences are not authoritative, or the objective good explains why trivial preferences lack authority)? I think the right answer,[2] appeals to the inapplicability of a central rationale for taking preferences seriously at all. Insofar as we care about preferences in moral and political theory, we care about them because they seem to capture what people *value*, or what expresses their own evaluative point of view. We believe that the fact that someone values something has normative or evaluative consequences: it *makes* the thing valued valuable, or *pro tanto* worth pursuing. But sometimes, as in cases of adaptation, preferences do not genuinely express what someone values. They do not, in Sumner's terms, "reflect the subject's own point of view" (Sumner 1996, 172). In other words, adaptive preferences display a failure of *autonomy*–a failure to express what *they* really value.[3] Though this idea is not precisely formed (an investigation into which, I should note, forms much of the remaining content of this essay), it seems right to say that when I adapt my preference to Columbia and away from Harvard, this is an adaptation *away* from my genuine values, or an attitude that accurately expresses my autonomous point of view. It is, as it were, putting on a kind of "mask": adopting a preference or evaluative attitude that does not reflect *me*. That adaptive preferences are non-autonomous seems essential to the concept and function of preference-adaptation. Adaptive preferences, after all, are *adaptations*:

---

[2] And, indeed, the most common answer. See, for instance, Sumner (1996, Ch. 6) and Elster (1981, 226-230).

[3] "Why are we reluctant to take at face value the life satisfaction reported by 'the hopeless beggar, the prevarious landless labourer, the dominated housewife, the hardened unemployed or the over-exhausted coolie'? …They do not lack enlightenment, or insight into the Platonic form of the good; they lack autonomy, …" (Sumner 1996, 166).

they alter the preferences we have in light of external circumstances, in particular, facts about the way the world is, or facts about what is available or unavailable to us (cf. Elster 1981). Adaptive preferences block one's own genuine evaluative attitudes insofar as maintaining our genuine attitudes, *given* facts about the way the world is, is worse; it is painful, frustrating, or otherwise disadvantageous. And hence it would appear that the essential fact of adaptive preferences, the fact that renders them a phenomenon *at all*, is that these preferences interrupt or mask our own genuine point of view. They are, for this reason, failures of autonomy.

There's another way to see this point. The normative failure of sadistic or shallow preferences is a matter of correspondence of a preference with some external measure, viz., moral demands, the requirements of respect for others, or some (independent of preference) measure of the objective good. But *adaptation* seems different: adaptive preferences do not (or do not necessarily) fail to conform to some external measure, but rather with an *internal* measure. But what is this internal measure? Surely adaptive preferences do not lack normative authority because they do not measure up to what I *previously* valued; this would render virtually all instances of changed preferences normative failures. Rather, adaptive preferences fail to measure up to the index of my genuine–autonomous–preferences or states of valuing. This seems to match up with, as it might be called, the phenomenology of such preferences. When we confront those whose preferences are adaptive, we have a tendency to think that their preferences do not really express what they genuinely value, or would value under conditions appropriate to developing autonomous preferences.

The connection between adaptation, autonomy, and normative authority forms the central question of this paper. In essence, it is this: though it is plausible to explain the normative failure of adaptive preferences *via* their lack of autonomy, is there any acceptable account of an autonomous preference that could form a principled method by which to distinguish preferences that lack normative authority as a result of adaptation? One way to investigate this question would be to argue for a particular account of autonomy, independently of that account's ability to offer a plausible explanation of the normative authority of preferences, and only then investigate whether autonomous preferences are normatively authoritative, or fail to be adaptive in a way that causes problems for ethics and social choice. Though this is perfectly open, this is not the style of argument I adopt here. Rather, I want to canvass a variety of potential accounts of the nature of autonomous preferences. I conclude that no such account is acceptable. By way of a conclusion, I assess what this verdict might mean for any plan to take preferences seriously in any normative domain.

3.4 Autonomy and Autonomous Preferences

As is noted by many theorists of autonomy, this concept is stretched thin. Nomy Arpaly identifies no less than *eight* concepts to which the term "autonomy" can and has been used to refer (Arpaly 2003, 118, Ch. 4 *passim*). Adding to the

difficulty here is that autonomy is generally a predicate applied not to preferences *per se*, but rather choices, decisions, or actions. For instance, we may say that an autonomous choice was a choice made without external interference, or a choice made on the basis of reasons (Sher 1997, 48-51). Thus the point at issue is what it might mean to ascribe autonomy to a *preference*. What might it mean to say, for instance, that my preference for Diet Coke over Diet Pepsi is autonomous? Or not autonomous? What *property* are we identifying?

I think accounts of autonomous preferences can be classed into roughly two categories. One might claim that the autonomy of a preference is an *historical* property: a property possessed by a particular preference in virtue of that preference's history or provenance; in particular, the way it was developed or instilled. Or it could be a *time-slice* property, a property that holds, or doesn't hold, of a particular preference at a particular time, regardless of that preference's history.

On the *historical* side, one might construe a preference as autonomous to the extent that it developed *via* the right sort of process. The project, then, must be to distinguish the right from the wrong processes. I can think of a number of potential accounts. First:

> *Historical Account One* (HA1): A preference is autonomous if and only if the agent in question engages in the process by which it is instilled or developed on the basis of reasons.

Second:

> *Historical Account Two* (HA2): A preference is autonomous if and only if the agent in question would endorse (perhaps under conditions of idealized reflection) the process by which it is instilled or developed.

Third:

> *Historical Account Three* (HA3): A preference is autonomous if and only if the process by which it is instilled or developed gives rise to preferences on the basis of reasons.

Fourth:

> *Historical Account Four* (HA4): A preference is autonomous if and only if the process by which it is instilled or developed is found on an "objective list" of right processes.

In addition to the historical accounts, there are a number of potential time-slice properties that could, in principle constitute the nature of of autonomous preferences. First:

> *Time-Slice Account One* (TA1): A preference is autonomous to the extent that this preference is endorsed by the agent in question (perhaps under conditions of idealized reflection).

Second:

> *Time-Slice Account Two* (TA2): A preference is autonomous to the extent that the agent in question maintains or possesses this preference on the basis of reasons.

Third:

> *Time-Slice Account Three* (TA3): A preference is autonomous to the extent a person has control over whether or not he or she maintains it.

Finally:

> *Time-Slice Account Four* (TA4): A preference is autonomous if and only if its object is endorsed under conditions of idealized reflection.

Just like the historical accounts, there could be additional time-slice accounts, but for the purposes of this paper, I'll limit my investigation to these.

I once again stress that I do not wish to get into a discussion of the exact nature of autonomy. Suffice it to say that each of these proposals seems to capture, at least in some way, the extent to which a particular preference expresses genuine states of valuing (away from which adaptive preferences are adaptations). To see this in more detail, take the historical accounts. Each historical account is tied together by a general thought that adaptation is, after all, a result of problematic *processes*: processes that "mask" an individual's genuine values.[4] Of course, what this amounts to is controversial, and is answered in different ways by each historical account. HA1-HA3 reflect the fact that many people think it plausible to believe that the concept of autonomy is tied very closely to the concept of acting for reasons, or at least acting for what one believes to be reasons. And so it may be plausible to say–as is reflected by HA1–that if a particular preference is the result of a process I engage as a result of reasons that I myself recognize (whether or not these are genuine reasons for action), this process generates autonomous preferences. To describe the right processes in this way is similar to some accounts of the nature of autonomous choice (cf. Sher 1997).

Similar thoughts motivate HA2. Let's say that I develop some preference for a particular object φ over some other object ψ, and that this preference was the result of some process P. But if we imagine, for instance, that I *endorse* P, perhaps for reasons, or perhaps under conditions of idealized reflection, then we might regard the preference in question as autonomous. (Notice that I construe the "endorsement" suggestion as potentially backward-looking. That is, a preference is autonomous to the extent that the process by which it was instilled is endorsed *at the time at which the preference is maintained*, whether or not it was endorsed at the time at which the relevant process was ongoing.[5]) It developed in a way that I regard as reflecting, say, *my own values*. HA2 differs from HA1, insofar as there

---

[4] This thought is well captured by Elster (1981, 226-8).

[5] Thanks to the editors of this volume for helpfully articulating this ambiguity.

is no constraint on HA2 that suggests the process must *actually* be engaged on the basis of reasons. It is enough to say, perhaps after the fact, that the process itself was valid, valuable, or otherwise endorsed by the person in question.

HA3 retains the broad connection between autonomous preferences and reasons, but draws this connection in a slightly different way. HA3 holds that a process is autonomous if that process itself includes coming to preferences *on the basis of reasons*. For instance, imagine that I prefer to refrain from smoking. If I came to that preference on the basis of becoming educated about the health effects of smoking, and thus saw reason to avoid smoking and developed preferences on this basis, this process is of the right sort and hence my preference is autonomous. But if my preference not to smoke was simply the product of disgusting images on the front of cigarette packages, which did not engage my capacity to recognize reasons, this process is not autonomous. HA3 guarantees that autonomous preferences are those that I develop on the basis of reasons I recognize. And if so, it seems plausible to say that such preferences will, broadly speaking, reflect my genuine conception of the good.

Finally, HA4 rejects the possibility of coming up with a single, unitary conception of the nature of processes of the "right sort". Akin to the "objective list" theory of well-being, which holds that a person's life goes well to the extent that it manifests particular items on a pre-determined list, this proposal says that processes generate autonomous preferences just in case such processes are identified on an objective list of autonomy-generating processes. Or, perhaps, that are *not* on an objective list of *wrong* processes; this list might include processes such as oppression, lack of opportunity, poverty, brainwashing, etc. Indeed, just this sort of view is floated by Sumner (in discussing the nature of autonomous life satisfaction). In despairing of the possibility of coming up with an adequate unifying account of the nature of autonomous processes, he (1996, 171) writes instead that:

> It appears, therefore, that neither of the currently dominant theories about the nature of autonomy is self-sufficient… However the details of a fully adequate view are worked out in the end, the implications for our theory of welfare are clear. Self-Assessments of happiness or life satisfaction are suspect (as measures of well-being) when there is good reason to suspect that they have been influenced by autonomy-subverting mechanisms of social conditioning, such as indoctrination, programming, brainwashing, role scripting, and the like.

Here Sumner relies on a list of processes that, according to Sumner, seem *not* to produce autonomous preferences. And, frankly, the list seems about right (whether or not there is an underlying theoretical unity to such processes). The processes noted by Sumner seem clearly to interrupt the extent to which a genuine preference can rightly be described as expressing *my* values.

Take now the time-slice accounts. TA1 also seems plausible strictly as a theory of autonomous preferences. I might maintain a particular preference for something

rather than another thing. But one thing that might indicate the extent to which this preference is autonomous is my *attitude* toward this very preference itself. For instance, imagine that I am a drug addict (cf. Lewis 2000, 70-71). I might prefer a dose of the drug to which I'm addicted rather than refraining from taking that dose. But is that preference autonomous? We won't know the answer to this question unless we know whether I endorse that preference, or take a pro-attitude toward it. (What *sort* of pro-attitude is proper for such endorsement is left unaddressed here, but, classically interpreted, it is a form of second-order desire or second-order preference, a "preference to prefer" the object of one's first-order preferences.) If I'm perfectly OK with my preference, despite the fact that I'm addicted, one might say that it is, in fact, autonomous: I endorse this preference (say, I "prefer to prefer it", or "desire to desire" taking the dose). Indeed, this account of the nature of autonomy forms the backbone of Frankfurt's influential account of the autonomous *will*, along with a number of influential accounts of the nature of personal value (Frankfurt 1988, 11-25; Lewis 2000; and Railton 2004, 43-68).

TA2 adapts a thought common to historical accounts specifically for time-slice accounts. As noted above, many believe that autonomous action is taken on the basis of reasons, at least reasons the person in question recognizes. But we might say the very same thing about autonomous preferences: when I *maintain* a particular preference for Diet Coke rather than Diet Pepsi *on the basis of reasons*, say, because the former tastes better, or because it contains fewer harmful chemicals, or because I like the color of the can, or because I find the advertisements less annoying, then we can say that the preference in question is autonomous. It is reflective of genuine values that I maintain.

TA3 reflects the generally intuitive thought that the nature of autonomy is closely connected to the nature of control. When I act autonomously, for instance, this seems to entail that my action was *self-authored*, rather than controlled by external forces. But we might put the same thought to work when it comes to autonomous preferences. If I have control over a preference, whether or not to maintain it or refuse to do so, it could be that when I maintain it, this preference is autonomous, just as an action or choice is autonomous to the extent that I have *control* over whether I perform the action in question or not.

Finally, TA4 takes a slightly different tack. A preference is autonomous, on this view, not to the extent that I endorse that preference, but to the extent that I endorse the *object* of that preference. One might think that this account is relatively thin, insofar as to prefer something just *is* to endorse that thing. But TA4 requires an additional condition: that one would endorse the object of the preference in question under a suitably specified set of idealized cognitive conditions. So, to take a simple case, it could be that I prefer some particular object just because I fail to maintain sufficient information about it (Haybron 2009, 185; Sumner 1996, 139). But I would not prefer that object, would not endorse the object of my preference, were I to maintain such information. In this case (depending on how one understands the nature of the idealized cognitive

conditions in question), the preference is not autonomous; the object of this preference would not be endorsed under idealized cognitive conditions.

It seems to me plausible, then, to say that each of these accounts connects with a strand of thinking about the nature of autonomy or the nature of autonomous preferences, specifically. Of course, insofar as these accounts focus on slightly different strands, they will not all be compatible. But the task of this paper is not to adjudicate between rival conceptions of autonomous preferences, but rather to determine whether *any* reasonable competitor accounts can adequately account for the normative authority of preferences, insofar as adaptation is a threat to such normative authority. I begin this investigation in the next section. One short note before I begin: this set of accounts of the autonomy of preferences is certainly not exhaustive. But, or so I shall argue in the conclusion, taking these accounts together allows us to draw an important conclusion about the relationship between adaptive preferences and *any* possible account of the autonomy of preferences.

3.5 HA1

According to HA1, autonomous preferences are developed by processes that are engaged in by people on the basis of reasons. But this cannot be the correct account of the normative authority of preferences. To see why, one need only note that adaptation to one's circumstances is a phenomenon that can occur for *perfectly good reasons*. Anyone, for instance, would see a reason to be more cautious around manholes if, in fact, one has recently fallen into one. But it also seems plausible to say that the mechanism of preferential adaptation can occur for similar reasons. When our preferences are frustrated, this is painful, it makes our lives worse than they might otherwise have been. And if there is no hope of fulfillment of those preferences, or if those preferences seem unlikely to be fulfilled, it may seem not just natural but positively *rational* to adapt our preferences to our circumstances (cf. Nussbaum 2000, 137-138).

But the rationality of the phenomenon of adaptation, when in fact it is rational, does nothing to vindicate the normative authority of preferences that undergo such processes of adaptation. To see why, consider the potential influence of preferences on theories of social choice. Note that a significant way in which preferences are adaptive are on the basis of existing social realities such as oppression, poverty, lack of opportunity. But adaptation to these realities can be rational for all the reasons just mentioned. But if and when policymakers are choosing to change social structures, or to assess the quality of such structures, it would be "ethically deeply mistaken" (Sen 1987, 46) to assess their quality in light of our rational preference revision to the social status quo.

I hasten to note that this problem also plagues a further account of the nature of autonomous preferences, viz., those preferences that are not formed as a result of *covert influence* (Colburn 2011, esp. 64-70). Covert influence is surely a feature of some instances of adaptation, but is not necessary. An individual's adaptation to his or her circumstances, like my adaptation to Columbia rather than Harvard, can

be perfectly up-front. Indeed, I might choose, for perfectly good reasons, to adapt in this very way. But this doesn't mean that there isn't anything "wrong" with adaptive preferences, as suggested by Ben Colburn. Their normative authority remains suspect even in light of the fact that they chose to develop them for perfectly up-front reasons. (Much of Colburn's discussion is motivated to distinguish adaptive preferences from "deliberate character planning," (Colburn 2011, 55; see also Elster 1981, 224) which is the "intentional shaping of desires" to, more or less, what one can get (Elster 1981, 224). But if we're concerned about the normative authority of preference, this is a distinction without a difference. As I understand it, adaptive preferences are characterized as *adaptations*; this adaptation can be conscious or unconscious, can take place "behind one's back", or as a result of deliberate choice. But this does nothing to alter the plausible judgment that such adaptations do not reflect *the real me*.)

Furthermore, however, there is good reason to believe that some normatively authoritative preferences will be the result of processes that I *do not* engage in on the basis of reasons. For instance, the fact that I grew up in certain formative years around fans of the estimable NFL franchise The Washington Redskins was a process that led me to prefer–even very deeply prefer–that the Redskins win, rather than that they lose. But it would be extremely implausible to say that this process is one that I undertook or engaged on the basis of reasons. This process, rather, *just happened to me*. But this preference is not adaptive and is surely normatively authoritative. If, for instance, the owner of the team wished to be genuinely beneficent to me, one way to do so would be to ensure that the team won more often, etc.[6]

If this is correct, it cannot be right to say that normatively authoritative preferences are autonomous *if* the right account of autonomous preferences is HA1. A process of adaptation can be engaged in on the basis of reasons, and indeed can be straightforwardly rational, even if those preferences clearly lack normative authority. In addition, straightforwardly normatively authoritative preferences can be the product of processes that were not engaged in on the basis of reasons.

3.6 HA2

According to HA2, autonomous preferences develop according to a process that the person whose preferences they are would endorse, perhaps after rounds of cognitive, idealized reflection.

HA2 can solve one of the problems that faces HA1. For instance, it could be that in considering the process that led to my preference that the Washington Redskins win, I am perfectly willing to endorse it. This process may well conform to my general values or, at the very least, would not alienate me to any meaningful degree. But notice that HA2 retains the first problem with HA1. Take some

[6] This is a problem for Colburn's account, as well. Cf. Bruckner (2011, 455-6).

process P, the end result of which is an adapted preference. Given the state of the world, or the state of my society, or some particular fact about me, it may be very clear that I cannot satisfy some preference of mine, or it may be good for me to develop some other preference more in line with the status quo. And if this is the case, I will have reason to engage P. But insofar as I have reason to engage P, it seems difficult to understand why I would regard P as a process unworthy of endorsement. After all, engaging it is rational given the state of the world, my society, or myself.

One might revise HA1 and HA2 in light of the problems noted here. One possibility is to say that an autonomous preference is one produced by a process engaged in–or endorsed–only for the *right kind of reasons*.[7] Note the shape of the reasons that lead me to adapt my preferences: such reasons are instrumental or strategic; they are not given my assessment of the *per se* or *intrinsic* value of the processes involved. Perhaps these are not reasons of the right kind. But what are? To salvage HA1 or HA2, any such account would have to rule out the possibility that the reason in question is strategic or instrumental in the way the cases of revision I've so far explored are. But it would have to do this without ruling out the normative authority of preferences we generally recognize as authoritative. But this is an extremely difficult problem. For instance, take one suggestion. Say that a reason based simply on the instrumental effects of a particular process is the wrong kind of reason to endorse, or to engage in, a process (at least for the purposes of determining preferential autonomy). This proposal seems essential in salvaging HA1 and HA2, but it cuts too deeply. Take a freshman entering college who has to decide whether or not to "go Greek", i.e., join a fraternity or sorority. Imagine that this person is neutral either way regarding the intrinsic benefits of going Greek, but nevertheless has some realization of the instrumental benefits of Greek-dom, viz., increased social connections, a leg-up in campus political races, etc. This person goes Greek for these reasons, and over time develops a strong preference for the Greek organization to which he or she belongs. In this case, the *process* that resulted in the preference was embarked upon for almost exclusively strategic or instrumental reasons. But we wouldn't say that the resulting preference for the Greek organization to which this person belongs lacks normative authority. In addition, it seems likely (or at least we can imagine that it is the case) that this person doesn't *endorse* the process by which he or she develops the preference in question (i.e., Rush week, or whatever other "loyalty building" exercises one engages in) for other than instrumental reasons. But the preference for his or her Greek house nevertheless maintains normative authority, and should be treated as such.[8]

But let's leave aside this point for the moment. There remains an important problem with HA1 and HA2 even if we are able to offer a plausible account of the "right kind" of reasons. It is this: to recognize a reason is to take some sort of positive attitude toward a state of affairs, fact, or other entity. But there is no

[7] For a discussion of the issue I note here, see Sobel (1994, 793 note 19).

[8] For a further discussion of the problems noted here, see Dorsey (2012, 412-415).

guarantee that *these positive attitudes* themselves are not the product of adaptation. I may very well recognize a *non-strategic* "right kind" of reason to, say, embark upon a process of adapting my preferences to my condition. But *this recognition itself* could be the product of problematic forms of adaptation; I might adaptively see a reason to remain in the midst of, rather than escape, poor or oppressive conditions and thus come to, non-strategically or for whatever "right kind" of reason, embark upon or endorse a particular preference-formation process *that further adapts my preferences to such conditions*. And thus insofar as the recognition of reasons itself is susceptible to the same failures of autonomy, we cannot look strictly to one's assessment–even "right kind of reason"-based assessment–of preference-formation processes for the sake of adequately ruling out adaptive preferences.

3.7 HA3[9]

HA3 is different. Rather than focusing on one's *endorsement* of, or *decision to undertake* a particular preference-formation process, it focuses on the content of the process itself, whether endorsed or not. Take the prototypical example of an insidious adaptive preference, i.e., a preference that one happens to have as the result of *brainwashing*. In this case, the problem with this preference seems to be the method by which it was developed, and the fact that this method is entirely antithetical to the autonomy or reasoning capacities of the agent in question. In other words, the process by which the preference was instilled was not one that instills a preference in a person *on the basis of reasons*. Rather, it instills a preference on the basis of, well, brainwashing. The same might be said for other examples of adaptation that are generally regarded as non-normatively authoritative. If I develop some preference simply as a result of oppression, I don't develop it, or so one can assume, on the basis of reasons.

But this account soon crumbles. Leaving aside the fact that it seems to succumb to the very same problems that characterize HA1 and HA2 (for instance, a process of adaptation *itself* could be a process that generates new preferences on the basis of reasons–even "right kind" reasons–given that the recognition of reasons can be colored by the process of adaptation), this account is under-inclusive. It is surely not the case that I developed my preference for Washington Redskins victories on the basis of anything like a *reason*; it just simply developed as a result of my social circumstances. No reasoning or engagement of my rational capacities were involved. This doesn't mean, however, that such a preference is not normatively authoritative.

Indeed, it would seem that *many* of the most hum-drum preferences we maintain are not the result of processes that instill preferences on the basis of reasons. My preference for black coffee over coffee with cream, for instance, was not the product of a reason-based process, but was developed, presumably, simply given the fact that my first experiences with coffee were without cream. My

[9] I'd like to thank Antti Kauppinen for excellent and thoughtful comments and conversation about this view.

preference for Beethoven rather than Mozart might be a result of music my parents played in the house when I was an infant. None of this is plausibly regarded as the development of a preference as the result of a reason-based process. And hence, it seems to me, even if HA3 were not over-inclusive, it is certainly under-inclusive and hence must be rejected.

3.8 HA4, and a General Argument against Historical Accounts

The final historical account eschews the possibility of finding any unifying or underlying feature of the "right" processes by which normatively authoritative preferences develop. This account instead simply settles for an "objective list", or a list of processes that can plausibly be said to generate autonomous preferences, and, thereby, a list of processes that can plausibly be said to generate non-autonomous preferences. Brainwashing will go on the latter list. Simple processes of preference formation while a child will, for instance, go on the former list. And so on.

One could critique HA4 for refusing to offer any further account of the rationale for inclusion of any particular process on the list of right or wrong processes.[10] But this would be to miss the point of the proposal. The proposal on offer is that there is no such rationale. And hence to complain on this basis would be to pound the table in favor of a more unifying account, an account HA4 denies the existence of. One may find it implausible to believe that there should be no more underlying unifier. But, as we have so far seen, it is difficult to find such a unifier that could plausibly explain the relative normative authority of preferences instilled by some processes rather than others. And so if one is committed to an historical account of autonomous preferences, then an historical account would seem to require abandonment of the hope for such a unifier.

But the problems with this view, I think, are shared by all historical accounts. First, it seems right to say that any particular preference formation process can produce preferences that are or are not problematically adaptive depending on *other* facts about the person in question. For instance, consider the example of preferences formed based on one's early social community, such as my preference for a Redskins touchdown. It seems right to say that *in some cases* this very process can also yield preferences that are problematically adaptive and hence non-normatively authoritative. For instance, consider the possibility that one's social condition is oppression- or poverty-ridden. If the individuals in my community, whose preferences are already adaptive, instill in me similarly adaptive preferences (say, preferences against social advancement, or preferences to remain in my poverty-stricken condition), few would say that this preference is normatively authoritative, or should be taken seriously in an assessment of my own good or of the success of social policies. And if this is correct, it would seem that a particular process of preference-development or installation is neutral with

[10] For a similar critique of the objective list theory of welfare, see Brink (2008, 32).

regard to the extent to which that preference is normatively authoritative. One might suggest instead that these examples display two distinguishable processes, viz., a process of *non-oppressive* community-based preference formation, versus a process of *oppressive* community-based preference formation. But this doesn't seem to help. Imagine that the oppressive character of my community exposed me, say, to the warmth of close and interdependent communities, which would not have occurred were my community not oppressed (a lack of oppression would have led to a stronger "rugged individualist" streak) (cf. Nussbaum 2000, 153-154). I may come to prefer this sort of closeness. But the process by which it was developed essentially involved the same oppression as my non-authoritative preference for destitution.

But what is the explanation of this? I think the most plausible one is that in the case of the Redskins, or in the case of communities with very close ties, there is no particular conflict between the preferences I have and any other attitudes I have. On reflection, I would *judge* a state of affairs in which the Redskins won (or in which my community was close-knit) a good one, or at least one to which I have an unproblematic attachment. But in the case of a preference for one's own poverty, it seems plausible to say *at first glance* that anyone considering, in the cold light of day, their own conditions as poor is likely to admit that the preference they maintain is maintained simply as a strategic device, or that given full information about one's self and one's circumstances, he or she would judge that that which one prefers is unattractive or lacks value. Given full information about the state of the world, it seems plausible to say, any person who maintains a preference for a life of poverty would be likely to judge that life undesirable in comparison to other lives; this needn't hold in the case of a preference for Redskins victories or a preference for close communities. And hence the problem of adaptive preferences is not trying to ferret out the processes by which these preferences were formed, but rather by coming to a kind of consistency or coherence among one's evaluative attitudes.

Here's another way to put this critique. The normative failure of adaptive preferences seems traceable to the fact that such preferences, broadly speaking, do not represent my genuine values; in this sense they are not autonomous. This helps to explain why my preference for Redskins victories is authoritative when my preference for conditions of squalor and poverty are not. But the extent to which a preference is reflective of *my values* seems more-or-less neutral with respect to the *processes* by which it developed.[11] Whether a preference is reflective of my genuine values has to do with that preference's connection to that person's wider evaluative attitudes. Thus even if we offer an account of the relevant preference-formation processes that eliminate all and only adaptive preferences, this account will not offer an account of autonomy that we have been seeking, viz., an *explanation* of our normative distrust of adaptive preferences. The processes by which adaptive preferences are developed seem to me evaluatively epiphenomenal. Instead, the right explanation is to be found in the extent to which

[11] For a more substantial argument on this point, see Bruckner (2009).

a preference does, or does not, capture my genuine values. And this is not an historical, but rather a *time-slice*, property.

3.9 TA1-TA3

Time-Slice accounts differ from historical accounts insofar as they identify autonomous preferences not on the basis of how these preferences came to be or the processes by which they were developed or instilled, but rather on the basis of facts about the relation between that preference and other psychological attitudes (perhaps counterfactual or idealized) of the person in question.

But this doesn't mean that time-slice accounts avoid the problems that plague historical accounts. Indeed, the first three seem to suffer from the same general problem that plagued the first three historical accounts. For instance, take TA1. TA1 holds that a particular preference is autonomous to the extent that it is endorsed, perhaps under idealized reflection. In this way, TA1 is similar to certain theories of the nature of the good. For instance, Peter Railton writes: "Let us then say that an individual's *intrinsic* good consists in attainment of what he would in idealized circumstances want to want for its own sake…were he to assume the place of his actual self" (Railton 2004, 54-55). David Lewis says something similar: "[A person] does not value what he desires, but rather he values what he desires to desire" (Lewis 2000, 70-71). Here it would appear that a particular preference or desire is *evaluatively* authoritative (i.e., its object is intrinsically valuable) to the extent that this desire is itself endorsed: someone *wants to want* something, or *desires to desire* it. Of course, a mere second-order desire isn't the only way one might construe the right sort of endorsement. Bruckner suggests that the relevant endorsement of a preference is an "all-in judgment that can conflict with a second-order preference" (Bruckner 2009, 317).

But this proposal gets into precisely the same trouble as HA2. Though it is possible that my idealized self is perfectly happy to endorse my preference for Redskins touchdowns, it is also very likely to be the case that this idealized self will endorse perfectly rational adaptive preferences given the various reasons one might have to develop them.[12] This is clear in, e.g., Bruckner's proposal. According to Bruckner, a gymnast who endorses her adaptive preference to

[12] One might try to revise TA1 in light of the problems noted here, and again identify a set of "right reasons" to endorse a particular preference. For instance, one might say that autonomous preferences are those that are endorsed *for their own sake*. But this proposal fails. It seems odd to say that my preference is autonomous only if I desire or endorse that preference *for its own sake*. I don't endorse a preference to be a philosophy instructor for its own sake; I endorse that preference because *I value being a philosophy instructor*, and hence having that preference is conducive to that which I value. But this doesn't mean that such a preference lacks normative authority or is problematically adaptive. There are other ways to construe this idea; regrettably I don't have the space to discuss them all, but I have done so elsewhere. I refer the reader to Dorsey (2012b, 413-415).

compete in regional tournaments rather than the Olympics, given her lack of ability to do so, would maintain a normatively authoritative preference. But this proposal is implausible. Though it may be rational[13] for her to maintain this preference, such a preference lacks the normative roles we assign to normatively authoritative preferences. We should not believe, in other words, that performing in regional competitions is better for this person than performing in the Olympics *would be*, despite her endorsed adaptation.

TA2 holds that a preference is autonomous insofar as the person whose preference it is maintains it on the basis of reasons. But as I have so far been at pains to argue, there are *reasons* to adapt one's preferences, and hence saying that autonomous preferences are those that individuals maintain or possess on the basis of reasons isn't going to solve the problem of normative authority. Though, perhaps, it remains a reasonable representation of one facet of the *autonomy* of preferences.

TA3 takes a slightly different tack with no better results. One might understand the notion of autonomy as picking out a form of control someone might have: autonomous actions are actions that, generally speaking, one has control over. Thus autonomous preferences are those over which one has control. If I can choose to maintain a preference or to rid myself of it, and I choose to maintain it, it would appear that this preference displays a form of autonomy in a perfectly respectable sense of that term.

Looking closely at this proposal, it is obvious that it cannot work. This is because the reasons to maintain adaptive preferences remain the same whether one has control over them or not. And so *if* one has control over whether to adapt to one's preferences or not, there will be a number of cases that render adaptation rationally justified. So far, time-slice accounts fare no better–and perhaps fare even worse–than historical accounts.

3.10 TA4

TA4 takes a substantially different approach than the previous accounts. This account states that autonomous preferences are for objects or states that are themselves endorsed or valued *under conditions* of cognitive idealization. Here's a reason to think that this might work. One plausible thought is that an individual's *genuine* values are those that he or she *would* develop if only the right sort of cognitive conditions held. Take, for instance, my adaptive preference for conditions of poverty or oppression. Though, as I actually am, I endorse this state,

[13] Indeed, Bruckner sometimes suggests that his proposal is simply that endorsed preferences are "rational". If this is correct, then I agree. But the tricky aspect of adaptive preferences is that their rationality does not entail normative authority. But if Bruckner wishes to claim that such preferences "ought to play the same role our rational deliberation as the rest of our preferences" (Bruckner 2009, 311), this seems incorrect for many cases he cites, for reasons discussed here.

it is hard to see how I could or would endorse or value that state if I adequately appreciated the status of alternative ways of living. If I fully understood the distinction between a life of poverty and a life free of destitution, and this distinction was made clear and vivid to me, it would be unlikely that I would keep such a preference (see, e.g., Brandt 1978, 113-126). Furthermore, it seems plausible that simple information about the world would not alter my general preference that the Redskins win the Super Bowl, as compared to some other football team, etc.

Notice that the "cognitive idealization" clause is essential for the proposal in question. Indeed, even a bare preference is itself a form of endorsement of a particular object. If I prefer to attend Columbia rather than attending Harvard, this is a form of endorsement of my attendance at Columbia. What is essential for this account, however, is the conjecture that cognitive idealization will slough off those preferences one has that are *merely* adaptive or do not express the genuine values such preferences are adaptations from.

Of course, this is a *big conjecture*. Is there any program of cognitive idealization that would guarantee that all and only non-adaptive preferences would be the result? I think a plausible answer begins by noting that a person's genuine values, as opposed to simply adaptive ones, must avoid being those that are *simply* a consequence of the world as it is. Elsewhere, I have argued that we can offer an account of an individual's preferences that is independent in this way if we combine a *full experience* constraint with a *coherence* constraint (Dorsey 2010, §4). My account of this view's virtues will be *very* sketchy, but take, for instance, my preference to go to Columbia rather than Harvard. This preference clearly depends on the contingent fact that I didn't get into Harvard. So what would my preference be if I examined the virtues of going to Harvard rather than Columbia *independently* of this fact? What if I fully understood and experienced both college careers, and revised my preferences in light of this experience toward a goal of coherence? Clearly, my adaptive preference would be revised away: in considering my options independently of my *actual* life, the "deep" values I maintain that point toward Harvard rather than Columbia will resurface, and (given their depth) will be maintained in any revision toward coherence.[14] The same seems correct about, say, adaptive preferences toward one's poor or oppressed conditions. As I come to possess full and adequate information about the various alternative styles of life I will develop potentially very deep preferences (preferences for greater freedom and opportunity, say) that are incoherent any adaptive preference toward poverty. If I have full experience of the alternative (even inaccessible) ways I might live, and my preferences are broadly coherent with what I find most important *given* such experience, we can and

[14] Much of this explanation relies on what it means to render a set of preferences coherent. See Dorsey (2010, 547-549). In essence, the general thought is that in cases of recalcitrant preferences, one revises based on a broadly Quinean rubric of "minimal mutilation", keeping fixed the "deepest" or most firmly held preferences, and/or those that would require more, rather than less, overall revision to one's preferential set.

should say that the preferences I maintain are not dependent upon any *particular* style of life or any particular facts about the world as it is. And hence such preferences cannot merely be adaptations to the way of life I maintain or the world around me. Rather, such preferences are reflective of what I *genuinely* value.

Much more needs to be said to defend this claim and so I won't put much weight on it here. However, even if there is a method by which to avoid cases of adaptation we've so far been discussing, there remains a problem with TA4. Let's call this the problem of "deep adaptation" (Dorsey 2012a, 100-108). No matter how much cognitive idealization one undertakes, if this form of cognitive idealization is really going to be rooted *in the person in question*, then it would appear that the only thing required for any rubric of cognitive idealization to retain adaptive preferences is for those preferences to be central to one's evaluative psychology. If my adaptive preference becomes entrenched, it seems hard to see how any preferences I might develop as a result of increased information or experience would themselves be strong enough or deep enough to override this deeply held adaptive preference in any bid for coherence (dorsey 2010, 551-552). But *however* the proper form of cognitive idealization is understood, there is no guarantee that adaptive preferences won't end up as some of the *deepest*–and most central–preferences one has.[15]

For instance, imagine that at some point in my life I come down with a chronic illness, say, type 1 diabetes. One method of coping with the disappointment of this illness is to strategically revise my preferences and evaluative attitudes: I come to hold that being diabetic is "central to my identity", say, and that monitoring my blood sugar manually is a form of "being in touch" with my body's chemical processes, etc., etc. This preference is adaptive. But the more I come to identify with it, the deeper it becomes, and the less likely it becomes that any plausible account of my genuine values will *not* include it. This result generalizes. Whatever program of cognitive idealization one prefers, it is possible, given the plasticity of human psychology, for an adaptive preference to pass the relevant idealization tests, and hence (according to TA4) remain normatively authoritative.

3.11 Conclusion

So what has been shown? So far no account of autonomous preferences has been able to adequately expunge adaptive preferences without also expunging perfectly respectable ones. But what conclusion are we to draw from this? Where are we to go from here? One possibility would be to offer some further account of the autonomy of preferences. But reflection on the failure of the accounts discussed here seems to offer a more general conclusion. As it happens, any plausible understanding of a person's true or genuine preferences or values is necessarily subject to problems of adaptation. Sometimes adaptive preferences, perhaps simply as a result of time, will form a crucial *part* of an individual's deepest

[15] For a further argument to this effect, see Hawkins (2008, 165-167).

self-understanding or point of view. (This was clear in the failure of TA4.) In short, adaptive preferences will, in many cases, form a significant aspect of our individual, autonomous, points of view. If so–necessarily–no adequate account of the autonomy of preferences will expunge all instances of adaptation.

In light of this, one might be tempted to shift the focus from the autonomy of a given preference, in determining its normative authority, to some other feature. In other words, if treating *autonomous* preferences as normatively authoritative cannot prevent adaptive preferences from possessing normative authority, we must focus on some other property that preferences possess or do not to explain the normative failure of adaptive preferences.

However, this strategy seems to me a poor fit for any view that seeks to take preferences seriously. In rejecting the normative authority of perfectly autonomous preferences, we lose a central motivation–explored above–for granting preferences a normative role. It seems plausible to say that preferences are important to take seriously, in coming to accounts of the nature of a person's good, or accounts of prudential or beneficent reasons, or accounts of proper social choice, because a person's preferences are supposed to represent something a person *values*. They are supposed to represent that person's way of evaluating the world around them, which itself is a significant normative fact. But if we reject the authority of preferences that, in fact, form a significant part of the way a person sees and evaluates the world around them, it's unclear to me why we should trust preferences to be normatively authoritative at all. Of course, if an individual's preferences are sadistic or shallow, or display various other *external* problems, we might downgrade them for this reason. But it's hard to see why, if a preference adequately corresponds to the various external measures (including morality, respect, objective value, etc.) *and* adequately corresponds to the relevant *internal* measure (i.e., the expression of genuine values), we should fail to grant it normative authority. That is, unless we refuse to accept the claim that motivated a normative interest in preferences in the first place.

I think there are two ways of addressing our current predicament. Call the first the "pessimistic" conclusion. This conclusion is motivated by the fact that to reject all adaptive preferences is to substantially weaken the motivation for taking preferences normatively seriously. The result of this conclusion is that preferences simply can no longer be trusted to play the normative role(s) to which they have been tasked: no plausible representation of *my* preferences can rule out all instances of adaptation. Perhaps the good, moral obligation, and social choice (or any other normative domain, for that matter) should look beyond my own values, no matter how genuine they are.[16]

For my money, the pessimistic conclusion would be a hard pill to swallow. Imagine, for instance, how strange it would be to deny that an individual's preferences have something to do with intrinsic value. It could be, for instance, that I strongly prefer a career as a professional baseball player to a career as a professional football player, just for its own sake. But it would appear that to deny the relevance of preferences would be to say that this fact can have nothing to do

[16] This conclusion is urged by Hawkins (2008), Nussbaum (2000), Sen (1987).

with the relative value of being a professional baseball versus football player for me. Any such value must be the outcome of facts that are independent of my own preferences. This is not incoherent to say, of course, but it is very implausible. Surely the fact that I prefer being a baseball player, other things being equal, makes it the case that being a baseball player is better for me than being a football player. This is not to say that facts beyond a person's preferences aren't relevant to value. But preferences surely *are* relevant in this way. Something similar should be said about social choice. Although there are plenty of arguments that seek to show that social choice should be neutral when it comes to conceptions of the good (Rawls 1993 and Quong 2010), surely if everything else is equal, making it the case that more rather than fewer of people's preferences are satisfied is something to be accepted in a social policy.

But the pessimistic conclusion is not the only option. Why believe that all adaptive preferences must lack normative authority? Especially if we believe that *problematic* forms of adaptation are those that "mask" an individual's genuine conception of the good, why must we say that adaptive preferences that *do not* mask such conceptions of the good fail to maintain normative authority? After all, "deep" adaptive preferences surely capture, rather than obfuscate, what I genuinely value–they plausibly represent *my* conception of the good. And so the "optimistic conclusion" is this: there is, in fact, a reasonable method by which to account for the distinction between normatively authoritative and non-normatively authoritative preferences. Assuming a proper round of cognitive idealization can be found–which I admit is a rather big assumption–we might say that it is TA4 (or something like it).[17] But this entails that "deep" adaptive preferences are normatively authoritative.

I think this is the right answer. But I will not argue for it here. It suffices to note that there is an important choice to be made in our treatment of the normative authority of preferences, especially adaptive preferences. Either we reject the normative authority of all adaptive preferences–and with it reject the motivation for granting normative authority to preferences at all–or we allow that some adaptive preferences, those that are suitably "deep", are normatively authoritative. Which choice to make is something I leave, at least for now, open.

References

Arpaly, N.: Unprincipled Virtue. Oxford University Press, Oxford (2003)
Brandt, R.: A Theory of the Good and the Right. Oxford University Press, Oxford (1978)
Brink, D.: The Significance of Desire. In: Shafer-Landau, R. (ed.) Oxford Studies in Metaethics, vol. 3, pp. 24–25. Oxford University Press, Oxford (2008)
Bruckner, D.W.: Colburn on Covert Influences. Utilitas 23, 451–457 (2011)
Bruckner, D.W.: In Defense of Adaptive Preferences. Philos Stud. 142, 307–324 (2009)
Colburn, C.: Autonomy and Adaptive Preferences. Utilitas 23, 64–70 (2011)

[17] Indeed, it seems to me that this is precisely the correct answer, and I have developed it in detail elsewhere. See Dorsey, (2012a, Ch. 3).

Dorsey, D.: The Basic Minimum: A Welfarist Approach. Cambridge University Press, Cambridge (2012a)
Dorsey, D.: Subjectivism without Desire. Philos Rev. 121, 407–442 (2012b)
Dorsey, D.: Preferences, Welfare, and the Status-Quo Bias. Aust. J. Philos 88, 535–554 (2010)
Elster, J.: Sour Grapes: Utilitarianism and the Genesis of Wants. In: Sen, A., Williams, B. (eds.) Utilitarianism and Beyond, pp. 226–230. Cambridge University Press, Cambridge (1981)
Frankfurt, H.: The Importance of What We Care About. Cambridge University Press, Cambridge (1988)
Hawkins, J.: Welfare, Autonomy, and the Horizon Problem. Utilitas 20, 165–167 (2008)
Kraut, R.: Desire and the Human Good. P. Addr. Amer. Philos Assoc. 68, 41–42 (1994)
Lewis, D.: Dispositional Theories of Value. In: Lewis, D. (ed.) Papers in Ethics and Social Philosophy, Cambridge University Press, Cambridge (2000)
Nussbaum, M.: Women and Human Development. Cambridge University Press, Cambridge (2000)
Quong, J.: Liberalism without Perfection. Oxford University Press, Oxford (2010)
Railton, P.: Facts and Values. In: Railton, P. (ed.) Facts, Values, and Norms. Cambridge University Press, Cambridge (2004)
Rawls, J.: Political Liberalism. Columbia University Press, New York (1993)
Rawls, J.: A Theory of Justice. Harvard University Press, Cambridge (1971)
Sen, A.: On Ethics and Economics. Blackwell, Oxford (1987)
Sher, G.: Beyond Neutrality: Perfectionism and Politics. Cambridge University Press, Cambridge (1997)
Sobel, D.: Full Information Theories of Well-Being. Ethics 104, 784–810 (1994)
Sumner, L.W.: Welfare, Happiness, and Ethics. Oxford University Press, Oxford (1996)

Chapter 4
"It Won't Be as Bad as You Think:" Autonomy and Adaptation to Disability

Jason Hanna

4.1 Introduction

People with disabilities often report enjoying a much higher quality of life than people without disabilities would expect. For instance, many people believe that their lives would be much worse if they were to become paraplegic. Indeed, some people believe that life with paraplegia would be barely worth living. Yet, perhaps surprisingly, many paraplegics report being roughly as happy as able-bodied people (see Gilbert 2006, 167, and Brickman, Coates, and Janoff-Bulman 1978). To take another example, people without colostomies would be willing to trade more months of life in order to avoid a colostomy than people with colostomies would be willing to trade in order to have their colostomies reversed (see Smith, Sherriff, Damschroder, Loewenstein and Ubel 2006).[1] This conflict in attitudes has come to be known as "the disability paradox." It has been found to arise for a range of disabilities and illnesses (Albrecht and Devlieger 1999).[2] There are three

Jason Hanna
Department of Philosophy
Northern Illinois University
DeKalb, IL
The United States
e-mail: jhanna@niu.edu

[1] For a brief summary of this and related findings, see Moller (2011).

[2] Although the label "disability paradox" has become fairly common, this terminology may be somewhat misleading, since the problem I will discuss applies to conditions that may not be appropriately regarded as disabilities. Further, the adaptive processes discussed below apply to areas of human life besides health. The following arguments thus do not rely on any precise account of disability; instead, for my purposes here, "disability" can be understood to comprise a wide range of health conditions that people typically want to avoid.

J. Räikkä and J. Varelius (Eds.): *Adaptation and Autonomy*, SAPERE 10, pp. 49–68.
DOI: 10.1007/978-3-642-38376-2_4 © Springer-Verlag Berlin Heidelberg 2013

possible responses to the paradox. First, it could be that people with certain disabilities overestimate their own well-being or quality of life. Second, it could be that non-disabled people underestimate the well-being or quality of life enjoyed by those with certain disabilities.[3] Third, it could be that some combination of these views is true: perhaps non-disabled people to some extent underestimate the quality of life enjoyed by people with disabilities, while people with disabilities to some extent overestimate their own quality of life.[4]

Many psychologists reject the first response and claim that the disability paradox is at least partly explained by affective forecasting errors. Someone makes an affective forecasting error when she makes a mistaken prediction about the effect of some event on her experiential quality of life (Gilbert and Wilson 2000). Affective forecasting errors are quite common. We often find that something we desperately wanted (or wanted to avoid) does not bring us as much satisfaction (or dissatisfaction) as we had thought. Affective forecasting errors have been extensively studied by contemporary psychologists, who have offered a number of explanations for their pervasiveness.[5] In the present context, one likely source of affective forecasting errors is the difficulty that non-disabled people have in recognizing how they might adapt to a disability. For instance, after someone develops a disability, her preferences and values may change, so that the disability frustrates her less than it otherwise would have. If an illness prevents someone from maintaining her grueling work schedule, she may come to appreciate the increased time that she now spends with her family, or the slower pace of her life, and so forth. But she may fail to anticipate in advance that she will adapt in these ways, and as a result the illness may seem worse in prospect than it would really be (see Gilbert 2006, 192-197).

Such affective forecasting errors raise moral problems that are especially pressing in the context of medical decision-making. For instance, when a recommended treatment would leave someone disabled, she may refuse on the grounds that she would rather live a shorter life without the disability than a longer life with the disability. Yet it may be that she would feel very differently after adapting to the condition in question. Likewise, someone may prefer one procedure to another on the grounds that it poses a lower risk of leaving her disabled. But her assessment of the procedures may be strongly influenced by her assessment of life with a disability.

What should we say about such cases? On the one hand, some ethicists have argued that affective forecasting errors compromise autonomy, so that decision-makers should take account of the likelihood that they will adapt. On the other, it may appear that decision-makers should have no regard for adaptive preferences

---

[3] For an argument to the effect that it is difficult to "explain away" someone's first-person assessment of her own well-being, see Barnes (2009).

[4] Of course, it could turn out that each explanation applies to some, but not all, of the cases in which the disability paradox arises.

[5] For useful overviews, see Gilbert and Wilson (2000), Wilson, Wheatley, Meyers, Gilbert and Axsom (2000).

they may form in the future. My central goal in this paper is to argue that both of these views are mistaken. If sound, my arguments suggest that legitimate moral concerns about a subject's well-being are sometimes misrepresented as concerns about her autonomy.

The paper proceeds as follows. I first explain why adaptive processes likely account for at least some of the preference changes that often follow the onset of disability (Section 4.2). I then consider two arguments purporting to show that affective forecasting errors compromise the autonomy of medical decision-making (Section 4.3). After challenging both arguments, I suggest that affective forecasting errors are problematic because of their effect on the decision-maker's well-being, not her autonomy (Section 4.4). Finally, I argue that adaptive preferences sometimes ought to be heeded and that, when they should not be, the reason may have little to do with autonomy (Section 4.5). This argument reinforces my claim that ethicists are often too quick to invoke autonomy in support of conclusions best defended through other means.

4.2 Adaptation to Disability

When people develop serious disabilities, they sometimes shift their priorities and change their values. As one group of psychologists observes, those who face restrictions in some areas of their lives may "find meaning and purpose in other aspects of their lives" and "even redefine, to themselves, what it means to be happy" (Ubel, Loewenstein, Schwartz, and Smith 2005, 61). For instance, in a survey of people with serious disabilities in the Chicago area, many respondents "reported how disability presented a new standard against which success is measured" (Albrecht and Devlieger 1999, 983). Consider the following statement from one respondent:

> Other people can't understand why I am so happy. They don't have the same appreciation of life. They would have to understand the satisfaction of using all my resources to conquer each day of challenges. (Albrecht and Devlieger 1999, 983)

Some people report that disability has brought them closer to others, reoriented their values, and helped them to see what is truly important in life:

> Disability was a huge shock for me. I always thought that old people were decrepit like my grandmother but never dreamed that it could happen to me, a 38 year old mother. I was bumbling through life, went to school, had kids and a job but took things for granted. Then, multiple sclerosis hit me like a lightening [sic] bolt– without warning. I went into a deep depression and was about to give up hope but I looked at my husband and kids and said: Get your head on straight. What's important here. [sic] My values were clarified by the shock. The big picture came into focus. My MS (multiple sclerosis) wasn't as important as my family and I wasn't going to let it ruin our lives. (Albrecht and Devlieger 1999, 983)

Such statements strongly suggest that, after becoming disabled, some people upgrade their assessment of life with a disability at least partly as a result of changes (or perhaps "clarifications") in their values or conceptions of happiness. A disability may prompt someone to reset or reorganize her priorities, and she may find that she is consequently able to derive satisfaction in ways that she had not fully anticipated.[6] Thus many people with disabilities report that while the experience of disability has changed their outlook on life, it has not worsened their life to the extent that they initially thought it would.[7]

In light of this point, it seems very likely that people without disabilities often make affective forecasting errors regarding disability. Further, it seems likely that these errors are at least sometimes explained by the failure to anticipate how their values and preferences would change in response to disability. The basic idea is this. When a person contemplates a disability in prospect, her evaluation will be influenced by the importance she places on activities or abilities that would be limited by the condition. If her values and preferences change so that these activities seem less important to her, then her assessment of life with the disability is likely to change, as well. Of course, this may not be the only source of affective forecasting errors regarding disability, but it is probably one source, and it is the one I shall focus on.

Even if someone's preferences change as a result of her disability, however, this may not show that her preferences are adaptive. In his classic discussion of "sour grapes," Jon Elster contrasts adaptive preferences with preferences formed through character planning (Elster 1983, 117-119).[8] Adaptive preference change, Elster argues, takes place "behind [one's] back" and is beyond one's deliberate control (Elster 1983, 117).[9] Character planning, by contrast, involves a deliberate effort to shape or regulate one's desires so that they can be more easily satisfied. The details of Elster's account may be challenged. In particular, it appears that many of our preferences are formed "behind our backs," without our exercising any control over their formation (Bruckner 2009, 316-317). On an alternative to Elster's view, a preference is adaptive when it has been subject to covert influence. A covert influence is one necessarily hidden from the subject, in the sense that the influence could not be the subject's reason for action if he were to become aware of it (Colburn 2011, 64-69). In LaFontaine's fable of the fox, it seems clear that the fox's newfound aversion to grapes is explained by the fact that the grapes are no longer available. But of course the fox cannot acknowledge

---

[6] Kahenman and Varey cite the following remark, made by a person with paraplegia during a television interview: "You probably think I am unhappy but you are wrong. And I used to think that I knew what suffering was, but I was wrong" (see Kahneman and Varey (1991, 144, note 6), cited in Savulescu (1994, 206)).

[7] Of course, this experience is not shared by all people who develop serious disabilities; some do not adapt very well. But many do.

[8] For an alternative way of drawing this contrast, see Bovens (1992).

[9] Although I here focus on adaptive preference *change*, the same thing is presumably true of adaptive preference *formation*.

this explanation – at least not if he honestly claims that it is the grapes' sour flavor that explains his aversion (Colburn 2011, 68).

It seems likely that, in at least some cases, post-disability preference change is likewise a product of covert influence or processes that necessarily take place behind one's back. Psychologists have drawn attention to the ways in which our "psychological immune systems" help us deal with apparent tragedy (Gilbert and Wilson 2000, 188-193; and Gilbert 2006, 177-178). When a person is a victim of severe misfortune– such as a severe disability or a natural disaster, say– she is likely to find ways to put a positive spin on her condition or find reasons to be happy with what she has. This is especially so when the condition is inescapable (Gilbert 2006, 201-204). As the psychologists Daniel Gilbert and Timothy Wilson note, the operation of the psychological immune system is "both powerful and invisible to the person it serves" (Gilbert and Wilson 2000, 188). If a former athlete claims that paraplegia has helped her to focus on what is truly important in life, we are likely to suspect the operation of coping mechanisms that downgrade activities or ways of life that may now be more difficult for her (or that upgrade activities or ways of life that are more readily available). But of course such a person will hardly explain her new perspective by claiming that she has simply changed her preferences so that they are easier to satisfy; indeed, she could not offer such an explanation if she claims that her preferences have changed because she now recognizes what is truly important in life. Of course, someone could engage in character planning and acknowledge that she has simply changed her preferences to fit whatever way of life is available to her. But especially in light of testimonies such as those above, it seems that many cases of adaptation to disability cannot be described in this way.

Further, at least some cases of adaptation to disability or illness illustrate some of the hallmark features of adaptive preference change. A study mentioned above found that current colostomy patients would be willing to trade far fewer months of life to have their colostomies reversed than former colostomy patients would be willing to trade to avoid having another colostomy. Former patients also rated normal bowel function as more important than did current patients (Smith, Sherriff, Damschroder, Loewenstein, and Ubel 2006, 691-693). This discrepancy in the assessment of life with a colostomy suggests the operation of adaptive processes. If current colostomy patients had changed their preferences as a result of character planning or because they had learned more about life with a colostomy, we should not expect these preferences to switch so soon after the reversal of the colostomy.[10]

[10] According to Elster (1983, 113), one salient difference between preference change through learning and preference change through adaptation is that the former cannot be reversed simply by a return to the original state. To be sure, the changes in preferences and values that we have been discussing may not always result in a clear preference reversal. In some cases the outcome of the adaptive process may be that the strength of the subject's preference has changed: her aversion to disability may simply be weakened after she has adapted. But such changes may still be adaptive.

My claim that the preference changes following disability are sometimes adaptive may be thought problematic. Some philosophers have worried that if we diagnose someone as having adaptive preferences, we may be led to dismiss her evaluation of her own quality of life. This may seem at best presumptuous and at worst disrespectful (Barnes 2009). In classifying a person's preference change as adaptive, however, I do not mean to imply that we can dismiss her first-person evaluation of her own life. Instead, I intend to be saying something about the process that led to the preference change. Even if I am correct that the preference changes following the onset of disability are sometimes adaptive in this sense, further argument is needed to show that we are entitled to discount or dismiss the assessment that people with disabilities make about their lives.[11] Moreover, to say that post-disability preference changes are sometimes adaptive is not to say anything about the sort of regard that we should have for these preferences. Indeed, as I shall argue below, many adaptive preferences are "worthy of pursuit" and merit the same sort of regard that other preferences do.[12]

4.3 Do Affective Forecasting Errors Compromise Autonomy?

Suppose that a patient makes a medical decision that is influenced by an affective forecasting error. (Here, and throughout, I focus only on affective forecasting errors that can be traced to a decision-maker's failure to anticipate how she would adapt to a disability.) Is such a decision autonomous? The answer to this question may be very important, since many biomedical ethicists hold that autonomous choices must be respected.

The philosophical literature seems to provide some support for each of two conflicting positions. The first position seems to be supported by a common view about adaptive preferences. It is often claimed that adaptive preferences are defective, typically on the grounds that they are non-autonomous. If this is so, then perhaps we should not have the same regard for our adaptive preferences that we should have for our other preferences. In most cases, we think that decision-makers have good reason to take account of preferences they will have in the future: if I know that I will want a nice retirement when I am older, it seems that I have reason to save for it now. But perhaps decision-makers should not take account of future adaptive preferences. For instance, it may be true that if a certain group were oppressed, they would over time come to prefer the social arrangement that results in their oppression. But this fact seems to provide no reason for them, or anyone else, to support the oppressive social arrangement. Likewise, it may be argued that

[11] To be sure, many philosophers think that we should be suspicious of first-person assessments of quality of life when the subject's preferences are adaptive. For instance, L. W. Sumner claims that we should regard self-reports of life satisfaction "as authoritative *unless there is evidence that they are non-autonomous*" (Sumner 1997, 171, quoted in Dorsey 2010, 542).

[12] For the quoted terminology, see Bruckner (2009).

when the preference changes following disability would be adaptive, a patient should not make her choice by considering these preferences.

A second position has been offered by biomedical ethicists concerned with affective forecasting. On their view, affective forecasting errors compromise autonomy, and a decision-maker's failure to recognize how she would adapt to a disability is likely to result in affective forecasting errors. Thus, a patient would be able to make an autonomous choice only if she takes account of her future adaptive preferences. To see the appeal of this view, consider an example. Suppose that an active young adult has been paralyzed from the waist down in an accident. He repeatedly requests, over a period of several days, that his doctor cease or withdraw life-sustaining treatment. The patient is fully informed of the medical facts and does not seem to be under the influence of any obvious judgment-distorting factors (such as intoxication or extreme pain). He simply insists that, for him, life in a wheelchair will be so impoverished that it will not be worth living.[13] Such a person believes that he will be miserable as a paraplegic, but he is far more likely to enjoy life than he thinks. On this basis, several philosophers have argued that his choice is non-autonomous. This view is expressed especially well by Rosamond Rhodes and James J. Strain (2008, 59):

> [P]hysicians must learn to be alert to instances when patients and families reach conclusions that are the effect of affective forecasting and recognize that, although those in its grip can speak and behave in ways that appear normal, choices that are generated by warped thinking should not be taken to be autonomous.

If such choices are non-autonomous, then the standard view in bioethics would hold that they need not be respected.

Each of the two views I have described should be resisted. I will first consider and reject arguments in favor of the second view, which holds that errors in affective forecasting compromise autonomy. Later, in Section 4.5, I will argue against the former view, which maintains that future adaptive preferences should be discounted or ignored.

4.3.1 The Ignorance Argument

Turn now to a first argument in favor of the view that affective forecasting errors compromise autonomy. According to this argument, such errors amount to a form of ignorance. It is generally acknowledged that ignorance can interfere with the exercise of autonomy. Suppose, for instance, that I refuse a recommended medical procedure because (and only because) I mistakenly believe that it would involve the amputation of my leg. Under these circumstances, there is good reason to claim that my choice is non-autonomous: as a result of ignorance, my choice does not really

[13] This case in some ways resembles the central example discussed in Savulescu (1994, 202-203).

express my values and preferences (see Feinberg 1986, Ch. 25). It may seem that affective forecasting errors can likewise result in ignorance of the sort that compromises autonomy. Immediately after someone has become disabled, she may falsely "believe that [her] current feelings of despair and anxiety will continue unabated into the future" or that she will no longer be able to enjoy life (Halpern and Arnold 2008, 1709). These false beliefs may well prevent the decision-maker from choosing in accordance with her values and preferences, and if so, her choice may appear to be non-autonomous. Call this "the ignorance argument." According to the ignorance argument, the paralyzed young adult of our earlier example has (probably) made a non-autonomous choice, since he (probably) has a mistaken belief about his long-term hedonic or affective reaction to disability.

While the ignorance argument may seem attractive, it faces serious difficulties. The central problem for the argument has to do with the link it would forge between autonomy and prudence. I will here consider a prudent medical decision to be one that best serves the patient's interests, and an imprudent medical decision to be one that fails to meet this criterion. At least in the medical context, it appears that imprudent choices are nearly always associated with the patient's failure to understand exactly what some possible outcome would be like for her. Recall that, according to the ignorance argument, knowing what some outcome would be like involves more than knowing the medical facts of the case. The patient must also make reasonably accurate predictions about her future affective or hedonic states, or about how she would feel under each outcome.[14] But if she had such knowledge, and she were committed to making a prudent choice (as most medical patients are), it is difficult to see how she could end up making an imprudent choice.

Here is another way to put the point. Suppose that a decision-maker has full knowledge of what each possible outcome would be like for her: she knows how satisfied she will be with each outcome, she knows how good (or bad) she will feel given each outcome, and so forth. (Perhaps we can imagine that it is as though she has already "lived through" each outcome and perfectly remembers what each was like.) In the context of medical decision-making, at least, it seems difficult to imagine a decision-maker who would choose an alternative that she knows will leave her less satisfied or feeling worse (or that can be reasonably expected to leave her less satisfied or feeling worse). And a medical decision that would leave the patient less satisfied is almost certain to be imprudent. If this is correct, then we can plausibly say of virtually anyone who has made an imprudent medical choice that she would have chosen differently if she'd had complete knowledge of what each alternative would be like for her.

[14] Rhodes and Strain (2008, 59), for instance, note that patients influenced by affective forecasting make "decisions based on distorted estimates of their future responses," and that this at least partly explains why their choices are non-autonomous. The main argument offered by Rhodes and Strain, however, seems more closely to resemble the appreciation argument, described in Section 3.3.2, than the ignorance argument described here.

Suppose, for instance, that Alice turns down a recommended medical procedure. Alice is fully informed about the medical facts of the case and undergoes careful and calm deliberation. She simply decides that the benefits of the procedure are not worth the risks. Now suppose that we strongly believe that the procedure is in Alice's interests, so that Alice has made an imprudent choice. If we are correct about this, then we will plausibly be able to claim that Alice has mistaken beliefs about what each alternative is like. (If she did not have mistaken beliefs about which alternative would leave her more satisfied, for instance, she presumably would not have made an imprudent choice.) And if such mistaken beliefs compromise autonomy, we will be able to conclude that Alice's choice is non-autonomous. The more general worry here is that if autonomous decision-making requires that the agent know what each alternative will be like in hedonic or affective terms, the result will be that virtually all imprudent medical choices are non-autonomous.[15] This result seems deeply problematic, especially if we claim to care about autonomy. (I describe further reasons for thinking that this result is troubling in Section 4.4, below.)

To be sure, some forms of ignorance compromise autonomy. If a patient has not been informed of the salient risks of some procedure, we may have good reason to suspect that her choice is non-autonomous. I have argued merely that we cannot make a corresponding claim when the patient is ignorant only of how happy she would feel under each alternative.

4.3.2 The Appreciation Argument

If I have been correct so far, then affective forecasting errors do not result in the sort of ignorance that compromises autonomy. But it may still be thought that such errors interfere with the exercise of autonomy for some other reason. Julian Savulescu claims that a patient's failure to anticipate how she will adapt to a disability can prevent her from vividly imagining the possible outcomes of her decision. Consider, for instance, a patient who has refused an amputation that is necessary to save her life. Savulescu argues that if this patient "had failed vividly to imagine what it would be like to adapt to life with a prosthetic leg, her previous desire to die would have been irrational and quite possibly obstructive [of her autonomy]" (Savulescu 1994, 210, emphasis added). Savulescu's point may not be that someone who has failed vividly to imagine possible outcomes is necessarily ignorant of some relevant information. Instead, his point seems to be that such a person is unlikely to appreciate her options (see also Rhodes and Strain 2008, 61). Autonomous decision-making requires more than mere information. It also requires that the decision-maker grasp the significance of relevant information so that she can discern how each alternative relates to her values and preferences. If a decision-maker fails vividly to imagine life with a disability in this way, there

[15] For a somewhat similar argument to the effect that all imprudent choices are impaired, see Hanna (2011). In the present context, I do not mean to endorse the claim that nearly all imprudent choices are non-autonomous. Instead, I have been arguing only that this result follows from the ignorance argument.

would appear to be a strong case that her choice is non-autonomous. Call this "the appreciation argument."

As applied to the cases we are considering, this argument would imply that people often assign low ratings to life with certain disabilities because they fail vividly to imagine such lives. It is not clear that this is true, however– at least not in all of the cases with which we are concerned. Return to our earlier example of the young man who refuses life-saving treatment because he does not want to survive as a paraplegic. The young man's problem may not be that he has failed vividly to imagine life with paraplegia. Instead, his problem may be that he fails to recognize how his evaluative perspective will likely change. From the perspective of his pre-adaptation values and preferences (viz., those he has prior to adapting to the disability), life with a disability may seem awful, even when it is vividly imagined. Indeed, there is some evidence that imagining certain disabilities in vivid detail may sometimes cause them to seem even worse to those without the disability (Ubel, Loewenstein, Schwartz and Smith 2005, 61). Yet from the perspective of the young man's post-adaptation values and preferences (viz., those he will have after adapting to the disability), the disability may not seem nearly as bad. The reason for this changed assessment may not be that he failed, at some time, vividly to imagine life with the disability. Instead, the reason may be, for instance, that he has changed his mind about how important it is to be able to walk, or continue his physically demanding job, and so forth.[16] If something like this explains the conflict between the pre-adaptation and post-adaptation assessment of disability, then the appreciation argument seems unconvincing.

At this point, a critic might argue that in order to appreciate "what it would be like to adapt" to a disability (as Savulescu puts it), someone must vividly imagine life with a disability from the perspective of the values and preferences she will likely have once she has adapted. We would then be able to claim that her autonomy is compromised if she is overly influenced by the more pessimistic pre-adaptation assessment of life with a disability. There are at least two problems for this suggestion. First, it seems to collapse into the ignorance argument. If someone were vividly to imagine an outcome from the perspective she will have if that outcome obtains, then she would presumably know how she will feel if that outcome obtains. The reply we are considering would then lead back to the problem discussed earlier in connection with the ignorance argument: there will be an overly tight connection between prudence and autonomy.

A second difficulty emerges when we reflect on the concept of autonomy. "Autonomy" may be a technical term among contemporary ethicists, and I cannot here consider the full range of views about what, precisely, autonomy is. Nonetheless, one core idea seems to be that a choice is autonomous to the extent

[16] Compare Moller (2011, 192-205). Moller argues that many people have attachments to particular activities and pursuits that may be impeded by a disability. On his view, this explains why they might be justified in, for instance, paying a great deal of money to avoid a disability, even though they recognize that they would not *feel* any less happy once they adapted to the disability.

that it reflects the chooser's own values (see Dworkin 1988, 31; and Savulescu 1994, 193-194).[17] Now the reply we are considering holds that if someone refuses a recommended procedure that would leave her disabled, her choice is autonomous only if it reflects the values she might hold in the future, after she has adapted to the disability.[18] But why should we think that autonomous decision-making requires one vividly to imagine outcomes from anything other than one's current values and preferences (see Scoccia 1990, 320)?

There may seem to be one promising way of linking a decision-maker's current values and preferences with those she might hold in the future. Most of us value our future well-being. And it seems plausible to suppose that my future well-being is in some way affected by how well my future preferences will be satisfied or my future values realized. If so, it may be argued that in order for me to achieve something I value now (namely, my future well-being), I must vividly imagine outcomes from the perspective of values and preferences I may hold later. Thus, in the paraplegia case, if the patient will obtain greater future preference satisfaction if he is given the life-saving treatment, we may conclude that his refusal of treatment is in tension with his current values. While this line of reasoning seems attractive, it leads back to the problems already discussed. Virtually everyone values his future well-being. And in the medical context, virtually every patient's primary goal is to improve his future well-being.[19] If we think that such patients

[17] Both Dworkin and Savulescu claim that one's autonomy is connected to one's ability to carry out a "life-plan." Dworkin focuses on the autonomy of persons, rather than the autonomy of choices, however.

[18] There is a further complication here. Return to our case of the young man who refuses a life-saving procedure that would leave him paraplegic. Suppose that the following things are true. If the young man is *not* given the treatment, he will not regret his choice: he will continue to prefer death to life with paraplegia. If the young man *is* given the treatment, he will come to prefer life with paraplegia to death. Which of these possible future perspectives should the young man use to evaluate his options? It seems arbitrary to say that he should evaluate them *only* from the perspective he will have if he is given the treatment. It may be argued that the young man must vividly imagine each outcome both from the perspective he will have if he is given the treatment and from the perspective he will have if he is not. But, *ex hypothesi*, these perspectives conflict. Defenders of the proposal we are considering may then need to find some way to adjudicate this conflict. I set this point aside in the following discussion, however, since I believe that the proposal faces even greater difficulties.

[19] There may be some exceptions. For instance, it may be that Jehovah's Witnesses who decline blood transfusions believe that they are sacrificing their own interests for the sake of a religious ideal. In fact, however, it is not clear that even this is typically the case. Jehovah's Witness publications on this issue focus largely on the bad medical outcomes that result from transfusion: "Each year thousands die as a result of transfusions; multitudes more get very sick and face long-term consequences. So, even from a physical standpoint, there is wisdom right now in heeding the Biblical command to 'abstain from blood.'" See "The Blood that Really Saves Lives," Watch Tower Bible and Tract Society of Pennsylvania, accessed November 23, 2012, `http://www.jw.org/en/publications/books/blood/the-blood-that-really-saves-lives/`.

choose autonomously only when their choices promote their future well-being, it will turn out, once again, that nearly all imprudent choices are non-autonomous.

4.4 Affective Forecasting, Autonomy, and Well-Being

It is worth briefly summarizing the dialectic so far. The ignorance argument and the appreciation argument purport to show that autonomy is compromised by affective forecasting errors, especially those associated with a decision-maker's failure to anticipate adaptation. Both arguments face a potentially serious problem. The ignorance argument seems to imply that virtually all imprudent choices (viz., choices that set back the patient's own interests) are non-autonomous. The appreciation argument either fails to establish its intended conclusion, or else has the same implication.

At this point, then, a critic may ask whether it would really be so bad for defenders of these arguments to admit that virtually all imprudent choices are non-autonomous (in the medical context, at least). After all, a central goal of health care is to promote patients' interests.[20] Further, medical decisions are likely to have serious, lasting, and irreversible effects on patients. For these reasons, it may be thought that autonomous medical decisions, especially life-and-death decisions, must meet a very high standard– a higher standard than would be required in other contexts (see Feinberg 1986, 117-124). On this basis, a critic may argue, it should not be especially surprising that nearly all imprudent medical choices are non-autonomous.

This response faces at least two problems. First, it may render the ignorance and appreciation arguments less attractive. These arguments may initially appear attractive at least partly because they seem to apply only under limited conditions. One can consistently place a premium on autonomy while recognizing that a fairly small subset of our choices, such as those induced by temporary impairment, are non-autonomous. On the other hand, if it turns out that the ignorance and appreciation arguments would classify nearly every imprudent choice as non-autonomous, they may turn out to be proverbial sheep in wolf's clothing. Further, the arguments would then fail to show that there is something uniquely problematic about choices that result from a decision-maker's failure to recognize how he will adapt to a disability; instead, such choices will be non-autonomous for the same reason that most imprudent medical choices are.

Second, in conceding that nearly all imprudent choices are non-autonomous, the imagined response trivializes the role of autonomy in biomedical ethics. The principle of respect for autonomy, which directs practitioners to respect autonomous patient decisions, is often contrasted with the principle of beneficence, which directs practitioners to act for patients' welfare (see Beauchamp and Childress 1994, Chs. 3, 5). The response we are considering, however, maintains that these two principles virtually never conflict (in actual

[20] This reply was suggested to me by the editors of this volume.

cases, at least). But if this is so, it is not clear that the principle of respect for autonomy plays any important role. What would be gained by appealing to the principle of respect for autonomy, as distinct from the principle of beneficence? Indeed, if prudence and autonomy are so closely linked, disputes about whether a choice is autonomous may typically reflect disputes about whether that choice is in the patient's best interests.

It may be thought that appeals to autonomy can still play an important rhetorical role. In particular, such appeals may serve to remind us that people are generally the best judges of their own interests, or that medical paternalism is typically counter-productive.[21] Yet even here, appeals to autonomy could have only a limited role. After all, defenders of the ignorance and appreciation arguments claim that those who fail to anticipate how they would adapt to a disability are prone to make imprudent choices, even when they "behave in ways that appear normal," are not impaired by mental illness or similar conditions, and are fully informed of the medical facts (Rhodes and Strain 2008, 59).

Fortunately, many biomedical ethicists concede that the principle of respect for autonomy sometimes conflicts with the principle of beneficence, and that imprudent medical choices may sometimes be autonomous (Beauchamp and Childress 1994, 271-291). Most claim that in such conflicts the principle of respect for autonomy nearly always prevails over the principle of beneficence. Call this the "trumping conception" of autonomy. Defenders of the trumping conception would claim that when a patient makes an autonomous choice to refuse an apparently beneficial life-saving treatment, her doctor may not perform the procedure against her will. Instead, they would argue, the doctor's role may be limited to providing relevant information and ensuring that the patient really is choosing autonomously. I suspect that the popularity of this view may at least partly explain why some philosophers have wanted to claim that autonomy is compromised by affective forecasting errors of the sort we have been considering.

Let me elaborate. Return to our example in which a young man refuses life-sustaining treatment because he believes that life with paraplegia will be miserable. If he falls unconscious before others are able to convince him of the contrary, it may appear that his doctors are permitted to proceed with the surgery, even over his prior refusal (Savulescu 1994, 210-211, and Rhodes and Strain 2008, 59-60). This verdict can be reconciled with the trumping conception only if we claim that the young man's refusal of treatment was non-autonomous (or at least that there is some reasonable likelihood that it was non-autonomous). And that, I suggest, is perhaps the main reason that some medical ethicists would be inclined to classify the choice as non-autonomous. More generally, when confronted with a patient's manifestly imprudent decision regarding medical treatment, defenders of the trumping conception have two options: they can either claim that the decision must be respected, or they can claim that it is

[21] For the view that medical paternalism typically is counter-productive, see Buchanan (1978).

non-autonomous. When intuition indicates that a choice does not merit respect, it may be tempting to classify it as non-autonomous.

The central concern here is that, in the absence of a fully developed account of autonomy, defenders of the trumping conception may classify choices as autonomous or non-autonomous simply in order to accommodate pre-theoretic intuitions about how the decision-maker ought to be treated. But obviously such a move is circular: if our conviction that a particular choice need not be respected is what makes it plausible to classify the choice as non-autonomous, then we can hardly appeal to the claim that the choice is non-autonomous in order to explain why it need not be respected.

We thus have good reason to explore alternatives to the trumping conception. The view that I prefer advances two claims. First, affective forecasting errors are problematic not because they produce non-autonomous choices, but rather because they are likely to produce imprudent choices. Second, the principle of beneficence can at least sometimes prevail over the principle of respect for autonomy, so that at least some imprudent but autonomous choices need not be respected. If this analysis is correct, then the arguments I have been criticizing mischaracterize legitimate moral concerns about well-being (e.g., that the patient's choice will harm him) as concerns about autonomy (e.g., that the patient's choice is not really his). For instance, on my view the doctor of the recently paralyzed man may be permitted to continue treatment. But this will be simply because such treatment is so strongly in the patient's interests (and perhaps because the consequences of the patient's choice would be irreversible).[22]

Some might argue that my treatment of this case is offensively paternalistic. A full discussion of this issue would take us too far afield. But there are a few points worth keeping in mind. First, many partisans of autonomy would support overriding the patient's choice in the sort of case we have been discussing anyway. Thus, the dispute is not so much about whether it is permissible to override the patient's choice as it is about why it is permissible to override the patient's choice. Further, my analysis of the case may force us to be more honest in recognizing that not all concerns about patient decision-making can be assimilated to concerns about autonomy.

4.5 When Are Adaptive Preferences Bad?

Earlier, I distinguished two different views that one might hold regarding the role adaptive preferences should play in patient decision-making. So far, I have been examining a view on which autonomous decision-making requires patients to take account of their future adaptive preferences. It is time now to consider the

[22] For the view that the irreversibility of consequences is relevant to the justifiability of paternalism, see Dworkin (1971, 122-123).

opposite view. On this view, decision-makers ought not to take full account of their future adaptive preferences.

As I noted earlier, this view may seem to receive support from the common claim that adaptive preferences are non-autonomous. Many philosophers have argued that because adaptive preferences are non-autonomous, they are "bad" (Colburn 2011, 68), "morally disvaluable" (Rickard 1995, 283) or otherwise suspect. At a minimum, I take it that if adaptive preferences are non-autonomous in a way that renders them bad or disvaluable, they should play a different role in deliberation than other preferences.

In some cases, this view seems plausible. Imagine that a woman is deciding whether to leave an abusive relationship. Suppose that as things stand, she is strongly inclined to leave the relationship because she does not want to put up with the abuse. But suppose she also has good reason to believe that if she stays in the relationship, her preferences will adapt so that she will eventually prefer the abusive relationship to the available alternatives. It seems that the woman should not heed these future adaptive preferences.[23] We might explain this judgment by claiming, more generally, that people should pay no heed to their future adaptive preferences when they are deciding what to do. Thus, the fact that the woman will be glad she chose to stay with her husband, or the fact that she will prefer the abusive relationship if she chooses to stay in it, does not provide good reason for her to make this choice. And this is so, it may be thought, simply because the relevant preferences are adaptive.

On inspection, however, this line of argument is overly simplistic. To be clear, the adaptive preferences in the abusive relationship case are problematic. Yet it seems perfectly rational for the recently paralyzed young man to make medical decisions in light of the fact that he will desire to live once he adapts to his condition. To take another medical example, suppose that someone's doctor has advised him that a colostomy could significantly extend his life. If the patient is like many of us, he may now prefer a shorter life to a longer life with a colostomy. If he were to have a colostomy, however, he would likely adapt to it and reverse his preferences (Smith, Sherriff, Damschroder, Loewenstein and Ubel 2006). And here it seems appropriate (or at least not obviously inappropriate) that this latter fact affect the patient's choice. The patient would not appear to be choosing

[23] It is worth noting, however, that the concerns frequently expressed about this sort of case may not be best interpreted as concerns about adaptive preferences. In some of the cases discussed by philosophers, a woman chooses to stay in a bad situation not because she has an adaptive preference for it, but rather because her options are so limited that the alternatives are even worse. Consider, e.g., Martha Nussbaum's (2000, 15-25, 140) description of two Indian women, Vasanti and Jayamma. For compelling criticism of Nussbaum's view that these women are subject to adaptive preferences, see H.E. Baber (2007, 111-114).

irrationally or non-autonomously if he consents to the colostomy on the grounds that he will later be glad that he did.[24]

What, then, distinguishes the cases in which it is problematic for decision-makers to display positive regard for their future adaptive preferences (such as the abusive relationship case) from the cases in which it is not problematic for decision-makers to display such regard (such as the paraplegia and colostomy cases)? In my view, the fact that some option will satisfy an adaptive preference does not give a decision-maker sufficient reason to choose it if doing so would leave her worse off on balance.[25] To illustrate the plausibility of this view, consider the case of LaFontaine's fox. Upon realizing that the grapes are out of reach, the fox claims that he does not want them because they are too sour. Instead, let us suppose, he prefers an apple. Now imagine that we offer the fox a choice between a delicious apple (which he now "adaptively" prefers) and some juicy grapes (which he preferred previously).[26] What should he choose? That depends. Suppose that although the fox prefers the apple, he would enjoy the grapes more than he would enjoy the apple. In this case, he does not have sufficient reason to choose the apple. Suppose, on the other hand, that the fox would enjoy the apple more than he would enjoy the grapes; perhaps his earlier experience of frustration really has changed his tastes in fruit.[27] In this case, the fox does have sufficient reason to choose the apple, even if his change in taste was adaptive. I fail to see why it would be irrational or non-autonomous for the fox to heed his adaptive preference when doing so will leave him better off.

---

[24] In a similar vein, Donald Bruckner (2009, 311) claims that "our adaptive preferences ought to play the same role in our rational deliberation as the rest of our preferences." On his view, whether *any* preference is normative or worthy of pursuit depends on whether it has been reflectively endorsed (Bruckner 2009, 317-319). The view I develop below does not embrace this reflective endorsement requirement.

[25] Martha Nussbaum (2000, 149) has argued that adaptive preferences do not count in the "social choice function" when they would downplay the importance of central capabilities needed for a good human life. Since my view in some respects resembles Nussbaum's, it is worth drawing two contrasts. First, I am not supposing that we should measure well-being in terms of capabilities; on my view, an adaptive preference might be problematic simply because it prevents one from satisfying more, or more important, preferences. Second, Nussbaum is focused on the role of preferences in social policy, while I am focused on the role that adaptive preferences should play in prudential decision-making.

[26] It might be thought that since the fox's preferences are adaptive, he would prefer the grapes if they were made available. In this case, however, we should question whether the fox ever had the adaptive preference for the apple; instead, it seems that he preferred grapes all along but simply deceived himself about this. See Baber (2007, 112).

[27] For an interpretation of the fable on which the fox changes his tastes rather than his beliefs, see Bovens (1992, 58).

Turn now to the case of the abusive relationship. Assuming that the woman has some alternative that would be better for her, she should not heed her future adaptive preferences: the fact that she will eventually prefer the abusive relationship if she chooses to stay in it does not provide a good reason for her to make this choice. (Of course, if the woman does not have a better alternative– if, for instance, the only alternative is life-threatening poverty– then perhaps she does have good reason to remain in the relationship. But her problem then would be limited options rather than adaptive preferences (Baber 2007).)

If all this is correct, then there appears to be nothing uniquely problematic about adaptive preferences that should disqualify them from consideration in prudential decision-making. In some cases, as in the disability cases described above, a decision-maker should heed her future adaptive preferences, because doing so will leave her better off than she would otherwise be. In other cases, where adaptive preferences should not be heeded, it is in virtue of a feature they share in common with some non-adaptive preferences: namely, that the decision-maker will be worse off for pursuing them.[28] It would be just as problematic for someone to pursue a desire to be oppressed or abused, even if it were not adaptive, since doing so would presumably leave him much worse off. On this view, when adaptive preferences are problematic, it is largely because of their relationship to the subject's well-being. This result would reinforce a conclusion reached in the previous section: we should sometimes appeal to the value of well-being in cases where many ethicists have appealed to the value of autonomy.

It is worth briefly considering one possible response to the argument of this section. A critic might argue that well-being and autonomy are closely related. In particular, it may be thought that because adaptive preferences are non-autonomous, the satisfaction of an adaptive preference contributes less to one's well-being than the satisfaction of an otherwise similar non-adaptive preference (see Dorsey 2010, 539). If so, then adaptive preferences will be problematic at least partly because of their relationship to the subject's autonomy. Even if it is in the subject's interests to satisfy a particular adaptive preference, the non-autonomous character of the preference would show that it counts for less in determinations of his well-being.

There are two points worth making in response to this objection. First, it is far from clear that non-autonomous preferences should be discounted in this way. To see why, compare two colostomy patients. Imagine that one has always preferred a longer life with a colostomy to a shorter life without one. Imagine that the second came to have this preference as a result of adaptive processes after the procedure. (Perhaps he originally agreed to the procedure only to make his spouse happy and thought that this choice would leave him worse off.) Let us suppose that, after the patients have had time to adapt, they are equally satisfied with their lives and

[28] For a similar point about the way in which a decision-maker should account for the fact that he will later be glad he made some choice, see Harman (2009).

equally strongly prefer a longer life with a colostomy. Is the second person's life going worse simply because his preference is adaptive (and thus, according to the view just described, non-autonomous)? That is far from clear.[29]

Second, the view I have offered in this section is consistent with the claim that non-autonomous preferences should be discounted for the purposes of determining someone's level of well-being. I have offered a view about the role that adaptive preferences, and especially future adaptive preferences, should play in prudential decision-making. On my view, an agent should heed his adaptive preferences when he will be better off if he does so. Even if this is correct, it may be that whether or not the preference is autonomous can help to determine how much its satisfaction would contribute to the agent's well-being. We can thus separate two questions. First: for the purposes of determining an agent's well-being, does it matter whether his preferences are adaptive? I am skeptical that is does, for the reasons described in the previous paragraph, but I avoid taking a strong stand on this issue. Second: when should an agent heed or pursue her future adaptive preferences? My view implies that she should do so when it would leave her better off, but this view is not committed to any particular account of well-being. Any plausible account of well-being would presumably imply, for instance, that it is in the recently paralyzed man's interests to continue treatment, at least if we assume that he will enjoy life after he adapts to his condition.

4.6 Conclusion

I have advanced two arguments regarding affective forecasting and adaptation to disability. First, I have argued that affective forecasting errors resulting from a decision-maker's failure to anticipate adaptation do not by themselves compromise autonomy. Second, I have argued that the case of disability shows that decision-makers sometimes ought to take account of future adaptive

[29] It may be thought that the second person's life is going worse if he does not identify with his adaptive preferences. On this view, if the second person were informed of the adaptive processes that led to the preference change, and he repudiated that change, then his life would be going worse than the first person's life. By contrast, if the second person were informed of the adaptive processes, and he then identified with the preference change, his life would be going as well as the first person's life. I am not convinced that this view is plausible. If a person has simply never considered whether his preferences are adaptive, it is not clear that we need to know which of these responses he *would* have in order to know how well he is actually doing. But even if I am mistaken about this, the proposal would show only that a preference does not count as much in determinations of well-being if the subject does not identify with it. We probably identify with some adaptive preferences and fail to identify with some non-adaptive preferences. Thus, the proposal may not point to anything uniquely problematic about adaptive preferences. See Bruckner (2009, 316-319). I am indebted to the editors of this volume for helpful discussion of this point.

preferences in prudential deliberation. Although I have advanced both arguments in the context of disability, they seem to have broader application. Thus, for the reasons given in Sections 4.3 and 4.4 above, it seems unlikely that the autonomy of a decision-maker's choice can ever be compromised merely by the fact that he has made a mistaken prediction of his future affective or hedonic states. And for the reasons given in Section 4.5, there may be a range of adaptive preferences that are worthy of pursuit and consideration.

Finally, the conclusions I have reached suggest that we should be cautious about appeals to autonomy in practical ethics. Although it may sometimes be permissible to interfere in choices induced by affective forecasting errors, and although some adaptive preferences may not be worthy of pursuit, this is not necessarily due to considerations of autonomy. More generally, I suspect that philosophers sometimes appeal to autonomy where appeals to other values would be more compelling. Recognition of this point may promote theoretical clarity. If we believe that autonomy is important, we should deploy it in some principled way. We must be clear about when a concern regarding a subject's choice or preference is best understood as a concern about autonomy, and when it is best understood as a concern about something else. Once this is done, we will be forced to confront more squarely the potential conflict between autonomy and other values, such as well-being. Our resolution to such conflicts will likely influence our views on a range of issues in practical ethics.

References

Albrecht, G.L., Devlieger, P.J.: The Disability Paradox: High Quality of Life Against All Odds. Soc. Sci. Med. 48, 977–988 (1999)

Baber, H.E.: Adaptive Preference. Soc. Theory Pract. 33, 105–126 (2007)

Barnes, E.: Disability and Adaptive Preference. Philos Perspect 23, 1–22 (2009)

Beauchamp, T.L., Childress, J.F.: Principles of Biomedical Ethics, 4th edn. Oxford University Press, Nwe York (1994)

Bovens, L.: Sour Grapes and Character Planning. J. Philos 89, 57–78 (1992)

Brickman, P., Coates, D., Janoff-Bulman, R.: Lottery Winners and Accident Victims: Is Happiness Relative? J. Pers. Soc. Psychol. 36, 917–927 (1978)

Bruckner, D.: In Defense of Adaptive Preferences. Philos Stud. 142, 307–324 (2009)

Buchanan, A.: Medical Paternalism. Philos Pub. Aff. 7, 371–390 (1978)

Colburn, B.: Autonomy and Adaptive Preferences. Utilitas 23, 52–71 (2011)

Dorsey, D.: Preferences, Welfare, and the Status Quo Bias. Aust. J. Philos 88, 535–554 (2010)

Dworkin, G.: The Theory and Practice of Autonomy. Cambridge University Press, New York (1988)

Dworkin, G.: Paternalism. In: Wasserstrom, R. (ed.) Morality and the Law, pp. 107–126. Wadsworth, Belmont (1971)

Elster, J.: Sour Grapes: Studies in the Subversion of Rationality. Cambridge University Press, New York (1983)

Feinberg, J.: The Moral Limits of the Criminal Law, vol. 3 Harm to Self. Oxford University Press, New York (1986)

Gilbert, D.: Stumbling on Happiness. Vintage, New York (2006)
Gilbert, D.T., Wilson, T.D.: Miswanting: Some Problems in the Forecasting of Future Affective States. In: Forgas, J.P. (ed.) Forgas JP (ed) Feeling and Thinking: The Role of Affect in Social Cognition, pp. 550–564. Cambridge University Press, New York (2000)
Halpern, J., Arnold, R.M.: Affective Forecasting: An Unrecognized Challenge in Making Serious Health Decisions. J. Gen. Intern. Med. 23, 1708–1712 (2008)
Hanna, J.: Paternalism and Impairment. Soc. Theory Pract. 37, 434–460 (2011)
Harman, E.: 'I'll Be Glad I Did It'. Philos Perspect 23, 177–199 (2009)
Kahneman, D., Varey, C.: Notes on the Psychology of Utility. In: Elster, J., Roemer, J.E. (eds.) Interpersonal Comparisons of Well-Being, pp. 127–163. Cambridge University Press, New York (1991)
Moller, D.: Wealth, Disability, and Happiness. Philos Pub. Aff. 39, 177–206 (2011)
Nussbaum, M.: Women and Human Development. Cambridge University Press, New York (2000)
Rhodes, R., Strain, J.J.: Affective Forecasting and its Implications for Medical Ethics. Camb Q Healthc Ethics 17, 54–65 (2008)
Rickard, M.: Sour Grapes, Adaptive Preferences and Objective Consequentialism. Philos Stud. 80, 279–303 (1995)
Savulescu, J.: Rational Desires and the Limitation of Life-sustaining Treatment. Bioethics 8, 191–222 (1994)
Scoccia, D.: Paternalism and Respect for Autonomy. Ethics 100, 318–334 (1990)
Smith, D.M., Sherriff, R.L., Damschroder, L., Loewenstein, G., Ubel, P.A.: Misremembering Colostomies? Former Patients Give Lower Utility Ratings Than Do Current Patients. Health Psychol. 25, 688–695 (2006)
Sumner, L.W.: Welfare, Happiness, and Ethics. Oxford University Press, New York (1997)
Ubel, P.A., Loewenstein, G., Schwarz, N., Smith, D.: Misimagining the Unimaginable: The Disability Paradox and Health Care Decision-Making. Health Psychol. 24(supp.), 57–62 (2005)
Wilson, T.D., Wheatley, T., Meyers, J.M., Gilbert, D.T., Axsom, D.: Focalism: A Source of Durability Bias in Affective Forecasting. J. Pers. Soc. Psychol. 78, 821–836 (2000)

Chapter 5
Autonomy and End of Life Decisions: A Paradox

Ben Colburn

5.1 Introduction

Suppose that we think it important that people have the chance to enjoy autonomous lives. An obvious corollary of this thought is that people should, if they want it, have control over the time and manner of their deaths, either ending their own lives, or by securing the help of others in doing so (see, e.g., Brock 1992, 11-12). So, generally, and even if we overall think that the practice should not be legalized on other grounds, it looks like common sense to think that considerations of autonomy tell at least somewhat in favour of legalizing at least some acts of suicide and voluntary euthanasia.[1]

In this paper, I argue for the opposite, paradoxical, conclusion. In fact, when we scrutinize the reasons for most end of life decisions, it turns out that they are seriously problematic from the point of view of autonomy. Full autonomy requires that we are responsible for the consequences of our decisions, and responsibility is precluded by non-voluntariness, which is to say decisions made because there are no acceptable alternatives. Since most end of life decisions are made for precisely this reason, it looks as though most such decisions are non-voluntary, and

Ben Colburn
Department of Philosophy
University of Glasgow
Glasgow
The United Kingdom
e-mail: Ben.Colburn@glasgow.ac.uk

[1] But see, e.g., Horton (2001) Keown (2002); unusually for opponents of voluntary euthanasia most of whom argue that some non-autonomy based consideration tells decisively against legalization, both Horton and Keown specifically deny that considerations of autonomy tell in favour of legalizing voluntary euthanasia.

J. Räikkä and J. Varelius (Eds.): *Adaptation and Autonomy*, SAPERE 10, pp. 69–80.
DOI: 10.1007/978-3-642-38376-2_5 © Springer-Verlag Berlin Heidelberg 2013

therefore undermine our autonomy: a discomforting and paradoxical claim. I argue that we should respond by taking the paradox to illuminate the context required by an autonomy-respecting framework for legalizing assisted suicide and euthanasia. People should have a legal right to a reasonable choice about when and how to die. However, this must go hand in hand with institutions that ensure, as far as possible, that such choices are made against a background which ensures, as far as possible, that people choose death clear-sightedly and not because nothing else is acceptable.

5.2 The Conditions of Autonomy

Autonomy, on the view that I am defending, is a global ideal, predicated of an entire life, which consists in deciding for oneself what is valuable and living one's life in accordance with that decision.[2] To be autonomous, on this view, means satisfying three conditions.

First, one must have the right sort of relation to one's ends and actions: that is, a disposition to *endorse* them. This might mean going through the actual process of reflecting on what one might want to do or be, and coming to a conscious decision about it; or it might mean the sort of wholehearted and unreflective pursuit of some goals or projects which makes it clear that one *would* consciously endorse them if the question arose. In respect of the end of life, this requires that the agent concerned either consciously decides to die – and so either commits suicide herself or requests assisted suicide – or possesses a disposition such that she would choose that if asked.[3]

Second, those endorsements must be *independent*: the ambitions and preferences that shape one's life must be one's own, if their pursuit is to contribute to the living of an autonomous life. My view is that independence consists in a property of the explanation of our decisions about what is valuable.[4] In particular, those explanations must not be covert, which is to say (qua true explanation) necessarily hidden from the agent concerned. That might be because it is being

[2] See Colburn (2010, Ch. 2) for a more extensive statement and defence of this way of construing autonomy. This conception of autonomy is similar to that defended by Joseph Raz (1986), and the ideal of individuality described by John Stuart Mill (1989).

[3] The difficulty of determining when an individual has such a disposition, if she doesn't actively exercise it, suggests that a prudent legal framework for end of life decisions would require an active expression of intent on the part of the individual concerned. Nevertheless, the underlying endorsing disposition is what does the normative work in such cases; active expressions of consent just give us evidence about that disposition's existence. This is important because it allows for the permissibility (morally, and perhaps legally) of non-voluntary euthanasia in the case of individuals who are unable explicitly to consent but of whom we have good reason to believe that their disposition under such circumstances would be to die.

[4] This is an analysis of independence developed in Colburn (2011, 52-71) and Colburn (2013).

kept hidden by some other agent: *you* say that the reason you endorse giving all your money away to a cult is because you've been touched by divine revelation, but we can see that the real explanation is that you were hypnotised by the cult leader, who has successfully kept the hypnotism secret from you. Or, it might be because the explanation is such that, if you were to become aware of it, you would cease to endorse the decision or action concerned. If the cult leader isn't as competent at covering her tracks as she had supposed, for example, you might find out about it – learning that *that* (rather than the greatness of Zeus) is the explanation for your decision to give away your money would very likely cause you to repudiate that decision. The crucial feature of such cases is that our decisions about what is valuable depend on something being withheld from us. Our decisions are independent only when this is not the case. In respect of end of life decisions, this requires that someone's decision to die must not be explained covertly, by some influence or cause which would cause her to decide otherwise if she knew about it.

These first two conditions of endorsement and independence elucidate what it means to decide for oneself what is valuable. For a life to be fully autonomous, on my conception, the individual concerned must also successfully pursue those ambitions, decisions and goals; or, at any rate, failure to satisfy an ambition diminishes one's autonomy to some extent.[5] That means, at least, that one's life goes in accordance with what one values. If someone wants to be a musician, and she is frustrated in that ambition, her life lacks autonomy for that reason. But one's life merely following the valued pattern won't be sufficient, because that could happen purely by chance, or because of benevolent external management. To capture the additional requirement that one *makes* ones life go in accordance with what is valuable, I incorporate a responsibility condition. Assuming that she endorses her decisions with independence, someone's life is autonomous to the extent that her life goes in accordance with those decisions, and she is responsible for the fact.

For full autonomy, she must be responsible in two ways.[6] First, she must bear *explanatory responsibility* for how her life goes. Her decisions must play a crucial causal role in what happens to her, and the shape of her life must be attributable to her agency. Second, she must be *evaluatively responsible*. This is a normative concept, which picks out a relation between an individual and a state of affairs

[5] In this I side with Joseph Raz's comments that "the autonomous person is one who makes his own life" (Raz 1986, 375), and against e.g. Gerald Dworkin, on whose view autonomy consists just in ones desires being structured a certain way, regardless of whether they are satisfied. See (Dworkin 1988, Chs. 1 & 2).

[6] For more on this distinction between explanatory and evaluative responsibility, see Colburn (2012). My distinction echoes those drawn by Thomas Scanlon between attributive and substantive responsibility, and by Ronald Dworkin between causal and consequential responsibility. To some extent, my theory is an attempt to improve on those important but somewhat ill-drawn distinctions. See Scanlon (1998, Ch. 6) and Dworkin (2000, 278-278).

which grounds some normative upshot, by which I mean some normative claims about what that individual must, may, or may not do or demand, which are true because the relation obtains, and all else equal wouldn't be true if it didn't. Evaluative responsibility is an important part of autonomy because living one's life is a matter not just of causing it to go a certain way, but also – to use Joseph Raz's phrase – of being *author* of it (Raz 1986, 369). Authority over one's life consists in standing in a normative relation to it which makes it appropriate to give one praise, blame, punishment, liability, reward and so on.[7] One is not the author of a life shaped by one's will being subverted in such a way that one cannot claim authorship of the consequences of one's actions. So, full autonomy requires both explanatory responsibility *and* evaluative responsibility in the sense which grounds such consequences.

In respect of end of life decisions, the condition of explanatory responsibility requires just that the individual's death is appropriately causally related to her decision to die, and that she is not prevented from choosing otherwise. Following Serena Olsaretti, I claim that the condition of evaluative responsibility requires, further, that her decision is *voluntary*. According to Olsaretti's theory, an act is voluntary so long as it isn't non-voluntary, and actions are non-voluntary when they are chosen because there are no acceptable alternatives (Olsaretti 2004, 139-141, 159-161). This involves both objective and subjective conditions being met. The standard of acceptability is here construed objectively and non-comparatively: an alternative is unacceptable if it involves falling below a certain absolute level (Olsaretti 2004, 154).[8] But it also matters what are the motivational reasons that explain one's action: one acts non-voluntarily if the *reason* one acts is to avoid unacceptable alternatives, so long as that is judged on an appropriate standard of acceptability.[9]

---

[7] David Owens has recently argued that certain duties and obligations are valuable *for us*, because being obliged is sometimes constitutive of relations we bear to others which are valuable; hence, there's even something valuable about being in a position where our receiving blame, guilt (and, we may add, punishment, though Owens doesn't) might be appropriate. My idea is similar: the value of autonomy consists in us having a certain type of relation not to other people but to the states of affairs that make up our lives. For Owens' proposal, see Owens (2012).

[8] Olsaretti, when fleshing out what standard she has in mind, says that an option is unacceptable "if pursuing or choosing it threatens some basic need." One might worry that positing an objective standard for what counts as unacceptable here runs counter to the underlying motivations behind a theory of autonomy, and that the combination of objective and subjective criteria mentioned here is therefore unstable. I tend to think not, because I think that it is possible to generate a standard by measuring unacceptable losses in autonomy; such a standard would pay heed to an individual's own perspective on her options, but need not treat an individual's judging an option to be unacceptable as either necessary or sufficient for the purposes of voluntariness. My thanks to Jason Park for pointing out the problem.

[9] There is further discussion of this conception of voluntariness, and of its connection with responsibility, in Colburn (2008, 101-111, 112-121).

Non-voluntariness vitiates evaluative responsibility because it involves the agent's will being subverted, either by circumstances or by the will of another. The consequences of those actions are alienated from the actor, and this alienation precludes the sort of relationship between her and those consequences which would make it appropriate to insist that she bears those consequences without compensation or aid. To illustrate, consider cases of coerced action. The problem with coercion is not that it eliminates free choice, strictly speaking. The highwayman's threat – "Your money or your life!" – does not prevent me from refusing to hand over the money, if I so choose. Indeed, far from precluding choice, its intelligibility and effectiveness presuppose my ability to choose. The point is that my choice is subverted by altering the character of my options: the highwayman arranges circumstances so that every practical alternative to surrendering my wallet, while remaining an option that I may choose, leads to a messy death. It is *this* that explains why my choice to hand over my money isn't sufficient to justify the exchange; or, to put it another way, leave me evaluatively responsible for my impoverishment in the sense that I have no claim against the highwayman for my money's return. One may not think that voluntariness is sufficient for evaluative responsibility; one might, for example, think that even some voluntary choices take place against such an unfair background that they vitiate responsibility. Still, so long as one thinks that there's something wrong with making people bear the consequences of choices which they perform as a result of the sort of force exemplified by coercion, one should think that voluntariness is necessary for evaluative responsibility.

To conclude, then, the theory of autonomy elaborated here suggests the following tests that we should apply to end of life decisions, to see whether they are compatible with the independence and responsibility components of individual autonomy. Both have to do with the explanation for an individual's decision to die. Those decisions lack independence to the extent that their explanations are covert; and they lack responsibility to the extent that those explanations hinge on the agent acting as they do because the alternatives are all unacceptable.

5.3 Decisions to Die

The preceding section concluded with some necessary conditions for autonomy which have to do with the explanation for why an agent chooses as she does: the explanation must not be covert, and it must not be that all alternative actions are unacceptable.

The first condition, of independence, will be failed in end of life decisions when those decisions fall short of the threshold of self-transparency required for independence. This happens, most notably, in cases of adaptation. Adaptive preference formation consists, in general, in the unconscious altering of our preferences in light of the range of options that we have available. In the context of end of life decisions, this is plainly a danger, given that such decisions are made at times of duress and stress. The danger is twofold. First, we might think there's a

particular danger of individuals deciding to die by dint of adapting their preferences to a perceived lack of alternatives; and second, the fact these decisions end in death makes it impossible subsequently to check the grounds of these decisions and to correct misapprehensions.

Despite this danger, and despite the high stakes, I think that adaptive preferences pose no peculiar problems here. There is no reason to believe that end of life decisions are generally adaptive in character: though some such decisions may lack independence, there is little reason to think that someone will wish to end her life only if she is deceived about the real explanation for her decision. Moreover, with a view to legislation, it will be comparatively easy to ensure that people's end of life decisions *are* independent, by making sure that they know about and reflect upon the reasons for their decision as far as possible, though (as Dudley Knowles points out) it is important to recognise the human costs of insisting upon people demonstrating their informed consent in situations which involve, as almost all end of life decisions will, significant emotional duress.[10] Still, in general, it seems to me that the problem of non-independence in end of life decisions is neither systemic nor difficult to address.

Nevertheless, there is a serious problem in the vicinity. This has to do with the second condition, of responsibility. As with adaptive preferences, the crucial determinant of responsibility is the explanation for an individual's decision. As the following argument shows, there is much more reason to think that these explanations will be malign in a way that affects responsibility, rather than independence.

1. A decision is non-voluntary if it is made because there are no acceptable alternatives.
2. If a decision is non-voluntary, the decider's evaluative responsibility (for the consequences of that decision) is undermined.
3. Anything that undermines evaluative responsibility undermines autonomy.
4. At present, most decisions to die are made because there are no acceptable alternatives.
5. (From 1 and 4) At present, most decisions to die are non-voluntary.
6. (From 2 and 5) At present, most decisions to die undermine the decider's evaluative responsibility.
7. Conclusion (From 3 and 6) **At present, most decisions to die undermine the decider's autonomy.**

The argument is valid. Moreover, we have good reason to accept each of the premises. Premise 1 is just a definition. Premises 2 and 3 follow from the theory of autonomy, responsibility and voluntariness defended in the preceding section. Premise 4 is borne out by studies of the reasons people seek to die: such decisions are almost always made in order to avoid alternatives correctly judged to be unacceptable. The usual reasons cited are the prospect of physical pain without

[10] I'm grateful to Dudley Knowles for this point; and see, for an illuminating discussion of the costs of demanding informed consent in an analogous situation, Knowles (2001).

prospect of relief; chronic dependency and loss of autonomy; and psychological trauma for the individual, friends, and family.[11] Those who act on these reasons aren't using an inappropriate standard for what counts as unacceptable: I have said little about what is necessary for an option to count as objectively unacceptable, but I take it that almost any account will have to judge that many of these paradigm factors do indeed count as such. So, Premise 4 is true, and the argument as a whole looks sound. This presents us with a paradox: far from expressing or supporting our autonomy, as common sense suggests, most end of life decisions in fact undermine the decider's autonomy.

There are two ways we might respond to this paradoxical conclusion. One possibility would be to treat it as a *reductio ad absurdum*. If the theory of autonomy set out above implies that almost all decisions to die diminish autonomy – including those made by competent individuals in full possession of all the relevant facts – then so much the worse for this theory of autonomy, or at least one of its elements. Our choice would then be to reject or revise either Premise 2 or Premise 3; that is, either the necessity of voluntariness for evaluative responsibility, or the importance of evaluative responsibility for autonomy.[12] As I indicated above, however, there are significant theoretical costs associated with rejecting either premise.[13] Taking this course also relies entirely on the uncritical acceptance of the intuition that there is indeed nothing troubling about most contemporary decisions to die. So, I conclude by showing what options we have if we retain both premises intact, and therefore accept the paradoxical conclusion.

5.4 The Circumstances of a Good Death

If the above argument is correct, it implies that as things stand most people's decisions to die detract from their autonomy, rather than (as common sense might suggest) either expressing or supporting it.

---

[11] For empirical data concerning the reasons people seek assisted suicide or euthanasia, see, e.g., Seale & Addington-Hall (1994), Emanuel, Fairclough, Daniels & Clarridge (1996), Mak, Elwyn & Finlay (2003), and Chapple, Ziebland, McPherson & Herxheimer (2006).

[12] My thanks to Robert Cowan for pushing this point particularly vigorously.

[13] For example, a critic might argue that, since the presence of dependency and pain are not intentionally caused by others, their being the factors that make alternatives unacceptable or unpalatable isn't enough to make these choices non-voluntary; the autonomous agent should be able to face these facts and come to terms with them. I think this criticism mistaken for two reasons. The first is that the painfulness and disempowerment of the alternatives to death clearly *are* caused by other humans, insofar as we have some (not unlimited, but considerable) means to relieve them through palliative medicine and social care, as I indicate below. The second, which I cannot really substantiate here, is that I anyway see no reason to distinguish between human-caused and non-human-caused reasons for unacceptability here: the crucial feature for voluntary choice is the nature of the chooser's options and the character of the choice she makes in light of them, not facts about the history of how those options come to have that nature. My thanks to Jukka Varelius for the objection.

One point worth clarifying is that this does not imply that we should prohibit either suicide or voluntary euthanasia. Even if there is a loss of autonomy involved in end of life decisions when those are made to avoid unacceptable alternatives, this would give us a reason to prevent people making and acting on such decisions only if that didn't involve a similar loss. Clearly, it would: someone prevented from acting in this way would be unable to shape the end of her life in accordance with what she judged to be valuable and appropriate, and her responsibility – both explanatory and evaluative – would be seriously compromised. This lose-lose situation is part of the paradox, I take it. Either we prevent people acting on their considered decisions, in which case they lose autonomy because their explanatory responsibility for their lives is undermined; or we don't, in which case their autonomy is undermined indirectly by dint of those decisions' non-voluntary character and the consequent effect on their evaluative responsibility.

We can break this deadlock by recognising two important points. The first is that it is not necessary that the reasons for most end of life decisions make them non-voluntary. People can, as a matter of fact, decide to die for reasons other than the unacceptability of their alternatives. If we can ensure that most people's end of life decisions are like that, then we escape the tragic situation posited by my earlier argument, by changing things so that Premise 4 is no longer true.

Sadly, looked at from the standpoint of institutional design, it is no easy matter to see how to use this insight. As I have shown, the status vis-à-vis autonomy of a given decision to die is a matter of the reasons on which the agent chooses or seeks to end her life. It is hard to design institutions and rules which detect and respond appropriately to such elusive internal features. One reason is the epistemic cost of doing so. Radically different motivations can issue in externally indistinguishable actions. Did the elderly parent who asked for an overdose of sedatives do so because she wanted to be out of pain, or because she was satisfied that her projects were done, or because she wanted her children not to see her in pain any more, or because she felt covert pressure from them to get herself out of the way? Designing a legal apparatus which reliable detects these differences is a formidable task. Moreover, even if these epistemic difficulties could be met, there would be more principled problems that might make us baulk, as I indicated in my earlier discussion of adaptive preferences: costs in terms of individuals' privacy, for example,[14] or the psychological costs of demanding that people bare their motivations and reasons at a painful time (Knowles 2001).

So, direct interference or scrutiny of individual end of life decisions looks neither practical nor desirable as a way to guarantee their voluntariness. The second step in breaking the tragic deadlock is therefore to identify another important tactic that social and legal institutions can deploy. All decisions take place against a background at least partly formed by such institutions. Merely by shaping that background in a certain way, those institutions can have an effect

[14] For arguments that privacy is important to individual autonomy, see, e.g., Henkin (1974), Reiman (1976), Kupfer (1987), and Cohen (2002).

upon whether decisions are voluntary or not, in a way that requires no direct interference with or scrutiny of those particular decisions.

To illustrate, consider the example (adapted from G.A Cohen) of someone who undertakes hazardous employment – let's say military service – because all alternatives involve starvation and penury for them and their family (Cohen 1988, 239-254). This individual acts non-voluntarily, as things stand. Now suppose, however, that this decision (whether or not to undertake military service) takes place against the background of a state guaranteed minimal income for all citizens.[15] That background changes the nature of our individual's options, such that the alternatives to military service are no longer unacceptable, for they no longer involve destitution. He or she may still choose to enter military service. However, given the background, we can now be confident that her motivation will be something else – the desire to see the world, a sense of justice or patriotism, sheer adrenaline-seeking bloodlust, or something like that – which preserves, rather than vitiates, the voluntariness of her decision.

Given this, I think the problematic reasons for most current end of life decisions gives us reasons to shape the context of these choices in an analogous way. We should shape the context of those decisions so that the explanations for people's decisions to die are, as little as possible, the unavailability of acceptable alternatives. In practice, this means two policies, each implemented with a view to neutralizing one of the two most important features which make the alternatives to death unacceptable: avoidance of pain, and avoidance of disempowerment and dependency.

The former implies an adequate regime of palliative care to ensure, as far as is medically possible, that individuals can enjoy surcease from pain even if they don't choose to die. This will ensure that avoiding unacceptable pain will not be the explanation for people's decisions to die in any but the most severe cases. It is striking that this argument reveals a deep and principled connection between legalizing euthanasia with the provision of palliative care and social support. In general, it is often assumed by opponents of legalized euthanasia that adequate palliative care is an alternative to legalized euthanasia: in particular, the possibility of the former undermines the argument for the latter.[16] My argument shows that this is not the case. Improvements in our ability to provide palliative care strengthen, rather than weaken, arguments to legalize euthanasia.

The latter is more complex. Institutional action to reduce such dependency might seem to be self-defeating, if it simply replaces one form of dependency and disempowerment with another. Recent research on the experience of adults with disabilities suggests that different mechanisms for external support result in wildly different experiences of dependency.[17] Identifying the features in virtue of which

[15] I have in mind here something substantial, like the basic guaranteed income argued for by Philippe van Parijs (1992, 1995).

[16] E.g. Childress (1998), Matthews (1998), Hamel (1998), the exchange between John Finnis and John Keown in Keown (1997, Chs. 1-6), Matersvedt et al (2003).

[17] For useful summaries of such research, see Macfarlane (1991), Zarb & Oliver (1993), Oliver & Barnes (2009), and Morris (2011).

this difference obtains will allow us to ensure that the alternatives to deciding to die are not unacceptably disempowering. We should, for example, extend the principles of reasonable adjustment to allow the very elderly to participate in social and civic life if they so wish, and ensure that the forms of support are shaped and agreed with by the individuals affected.[18]

Much more work is required on the details of both these proposals, and it is important to recognise that their effect will only ever be mitigative: it will be impossible to eliminate all factors which made some individuals' decisions to die non-voluntary. However, they offer a model for how we might provide institutional support which allows us to deal with the uncomfortable conclusion of Section 2: even if we accept the conclusion that (as things stand) many decisions to die diminish our autonomy, it is open to us to deploy our social and legal institutions to eliminate, as far as possible, the toxic background that makes those decisions non-voluntary. Under such circumstances, we can be much more confident that when people *do* nevertheless decide to die, they are doing so voluntarily, and hence in a way that expresses, rather than diminishes, their autonomy.

5.5 Conclusion

I have argued that end of life decisions pose a serious problem for autonomy. Whether or not a decision contributes to your autonomy depends on whether it is voluntary, and – at least as things currently stand – there is good reason to believe that most end of life decisions are non-voluntary, because they are made to avoid alternatives correctly perceived as unacceptable. If the reader thinks that this conclusion is absurd – because there plainly is no autonomy-based problem with the majority of contemporary decisions to die – then they will take this argument to be a *reductio* of the theory of autonomy defended at the start of this paper. I suggest a different interpretation. This argument genuinely reveals a troubling problem for many current end of life decisions. To solve it, we must change the circumstances of those decisions so that, as far as possible, the individuals concerned always have acceptable alternatives. People are indeed deeply vulnerable when they are contemplating decisions to die, because their chances of autonomously navigating the final most important event in their lives are at risk. The measures above indicate a way in which we can mitigate that vulnerability, and offer people the best possible chance of an autonomous death.[19]

---

[18] Jenny Morris (2011, 16-17) identifies these as two of the key factors which disabled adults regard as making the difference between support which empowers and enables, and support which disempowers and creates dependency.

[19] Thanks to Jukka Varelius and Juha Räikkä for their helpful comments on a draft of this paper. Thanks also to Robert Cowan, Jennifer Corns, Derek Brown, Adam Rieger, Emily Askham, Dudley Knowles, Gareth Young, Jason Park, and Harriet Morse, for discussion of the arguments it contains; and to Daniel Elstein, Hallvard Lillehammer, Serena Olsaretti and Jonathan Quong for discussions of different elements of the underlying theory.

References

Brock, D.A.: Voluntary Active Euthanasia. Hastings Cent. Rep. 22, 10–22 (1992)
Chapple, A., Ziebland, S., McPherson, A., Herxheimer, A.: What People Close to Death Say about Euthanasia and Assisted Suicide: A Qualitative Study. J. Med. Ethics 32, 706–710 (2006)
Childress, J.F.: Religious Viewpoints. In: Emanuel, L.L. (ed.) Regulating How We Die: The Ethical, Medical and Legal issues Surrounding Physician-Assisted Suicide, pp. 120–147. Harvard University Press, Cambridge (1998)
Cohen, G.A.: Are Disadvantaged Workers Who Taker Hazardous Jobs Forced to Take Hazardous Jobs? In: Cohen, G.A. (ed.) History, Labour and Freedom, pp. 239–254. Clarendon Press, Oxford (1988)
Cohen, J.: Regulating Intimacy: A New Legal Paradigm. Princeton University Press, Princeton (2002)
Colburn, B.: Authenticity. In: Levey, G. (ed.) Autonomy, Authenticity, Culture (forthcoming, 2013)
Colburn, B.: In Defence of Comprehensive Liberalism. Philos Pub. Issues 2, 17–29 (2012)
Colburn, B.: Autonomy and Adaptive Preferences. Utilitas 23, 52–71 (2011)
Colburn, B.: Autonomy and Liberalism. Routledge, New York (2010)
Colburn, B.: Debate: The Concept of Voluntariness. J. Pol. Philos 16, 101–111 (2008)
Dworkin, D.: Sovereign Virtue. Harvard University Press, Cambridge (2000)
Dworkin, G.: The Theory and Practice of Autonomy. Cambridge University Press, Cambridge (1988)
Emanuel, E., Fairclough, D., Daniels, E., Clarridge, B.: Euthanasia and Physician-Assisted Suicide: Attitudes and Experiences of Oncology Patients, Oncologists, and the Public. Lancet 347, 1805–1810 (1996)
Hamel, R.: A Better Approach to Care of the Dying: Catholic Healthcare and the Catholic Community Can Present an Alternative to Physician-Assisted Suicide. Health Prog 79, 54–59 (1998)
Henkin, L.: Privacy and Autonomy. Colum L. Rev. 74, 1410–1433 (1974)
Horton, R.: Euthanasia and Assisted Suicide: What Does the Dutch Vote Mean? Lancet 357, 1221–1222 (2001)
Keown, J.: Euthanasia, Ethics and Public Policy. Cambridge University Press, New York (2002)
Keown, J. (ed.): Euthanasia Examined. Cambridge University Press, Cambridge (1997)
Knowles, D.: Parents' Consent to the Post-Mortem Removal and Retention of Organs. J. Appl. Philos 18, 215–227 (2001)
Kupfer, J.: Privacy, Autonomy and Self-Concept. Am. Philos Quart. 24, 81–89 (1987)
Macfarlane, A.: Ageing and Disability. In: Laurie, L. (ed.) Building our Lives: Housing, Independent Living and Disabled People, Independent Living and Disabled People, Shelter, London (1991)
Mak, Y.Y.W., Elwyn, G., Finlay, I.G.: Patients' Voices are Needed in Debates on Euthanasia. BMJ 327, 213–215 (2003)
Matersvedt, L.J., et al.: Euthanasia and Physician-Assisted Suicide: A View from an EAPC Ethics Task Force. Palliative Med. 17, 97–101 (2003)
Matthews, H.: Better Palliative Care Could Cut Euthanasia. BMJ 317, 1613 (1998)
Mill, J.S.: On Liberty. In: Collini, S. (ed.) On Liberty and Other Writings, pp. 1–116. Cambridge University Press, Cambridge (1989)

Morris, J.: Rethinking Disability Policy. Joseph Rowntree Foundation, London (2011)
Oliver, M., Barnes, C.: Disabling or Enabling Welfare: What Next for Disabled People. In: Oliver, M. (ed.) Understanding Disability, 2nd edn., pp. 119–133. Palgrave, Tavistock (2009)
Olsaretti, S.: The Concept of Voluntariness – A Reply. J. Pol. Philos 16, 112–121 (2008)
Olsaretti, S.: Liberty, Desert, and The Market. Cambridge University Press, Cambridge (2004)
Owens, D.: Shaping the Normative Landscape. Oxford University Press, Oxford (2012)
Parfit, D.: Reasons and Persons. Oxford University Press, Oxford (1984)
van Parijs, P.: Real Freedom for All. Oxford University Press, New York (1995)
van Parijs, P.: Arguing for Basic Income: Philosophical Arguments for a Radical Reform. Verso, London (1992)
Raz, J.: The Morality of Freedom. Clarendon Press, Oxford (1986)
Reiman, J.: Privacy, Intimacy, and Personhood. Philos Public Aff. 6, 26–44 (1976)
Scanlon, T.: What We Owe to Each Other. Harvard University Press, Cambridge MA (1998)
Seale, C., Addington-Hall, J.: Euthanasia: Why People Want to Die Earlier. Soc. Sci. Med. 39, 647–654 (1994)
Zarb, G., Oliver, M.: Ageing with a Disability: What do They Expect After All These Years? University of Greenwich, London (1993)

Chapter 6
Gendered Adaptive Preferences, Autonomy, and End of Life Decisions

Serene J. Khader

6.1 Introduction

Other essays in this volume focus on how we may adapt our preferences to available health opportunities. In the late stages of a terminal illness, we may come to underestimate the desirability of a high-functioning life. Increases in the availability of life-extending technology may make it difficult to perceive value in the normal experience of aging. However, health opportunities are not the only circumstantial factors that stand to influence our end of life choices. End of life decisions call up questions about what we believe death is, how we want others to see us, and what we think we deserve from them – among other deep questions about value. Biological limitations and the availability of medical technologies are not the only types of external factors that influence our views about these questions of value. Our individual answers to such questions are related to the religious views to which we have been exposed, the strength of our familial and social ties, the occupational and other life opportunities that our society has made available to us, and the views our society has promoted about our worth and the worth of people like us. A lifetime of social limitations (and/or opportunities) has shaped the perspective from which any of us decides whether to pursue life-prolonging treatment, palliative care, or death.

In this essay, I am interested in how a specific set of situational limitations affect our end of life decisions – limitations imposed by *unjust social arrangements.* Much has been written about how conditions of social injustice affect people's

Serene J. Khader
Department of Philosophy
Stony Brook University
Stony Brook, NY
The United States
e-mail: Serene.Khader@sunysb.edu

J. Räikkä and J. Varelius (Eds.): *Adaptation and Autonomy*, SAPERE 10, pp. 81–100.
DOI: 10.1007/978-3-642-38376-2_6 © Springer-Verlag Berlin Heidelberg 2013

views of what is possible for them (Nussbaum 2001, Khader 2011, Superson 2005, Babbitt 1993, Sen 1990). There is also a large literature on how dynamics of privilege and marginalization can impact people's end of life choices. For instance, some argue that African-Americans may be less likely to opt for palliative care because of a lack of trust in the medical establishment (Crawley and Payne 2000). There is also concern that people with disabilities will prefer death because of a lack of social supports that would allow them to flourish (Ho 2008, 203). Critics of assisted suicide point out that women are overrepresented among assisted suicide patients – and that many of the women in the Dr. Kevorkian cases were not even terminally ill (Solomon and Noll 2008). Women also seem more likely than men to believe that their doctors want them to undergo palliative care (Johnson, Lin, et. al. 2000). Men seem to be particularly unable to handle the thought of being dependent on others (Hayes, Lee, et. al 2011) and some speculate that this fact explains men's preference for life-sustaining treatments over the intimacy of palliative care (Rietjens, Descheppet, et. al 2012, 1286).[1]

I will ask here whether and when it is appropriate for medical practitioners to intervene in end of life choices that appear adapted to unjust circumstances. I will focus on gendered preferences related to the giving and receiving of care. My answer will be that there are autonomy-based reasons for intervention when the preferences in question are lower-order preferences that are a) either based on a misunderstanding of the non-normative facts or b) clearly inconsistent with the rest of the patient's values. Lower-order preferences here are preferences for specific courses of action – such as preferences for certain courses of treatment. I distinguish these from the higher-order values and beliefs that constitute the perspective from which a person evaluates potential courses of action. But lower-order preferences inconsistent with the facts or the patient's values are only a small subset of preferences that can reasonably be called "adapted to unjust social conditions." Many sorts of preferences that are adapted to injustice – though they produce tragedy during the lifetimes of those who have them – are not appropriate to question or override at the end of life. Some preferences that are adapted to injustice are harmful to the agent over the course of a lifetime; I have elsewhere called these "inappropriately adaptive preferences" or IAPs (Khader 2011). The reason that medical professionals should not intervene to change higher-order IAPs at the end of life turns out be that higher-order IAPs can be autonomous, and that the promotion of perfectionist values is not an appropriate goal at the end of life.

I begin by explaining that higher-order preferences that are adapted to injustice are typically morally problematic because of perfectionist concerns and concerns about justice these are distinct from autonomy concerns. As I argue in the first section of the chapter and have argued elsewhere, plenty of IAPs pass tests for procedural autonomy, and we see them as morally problematic nonetheless

[1] Bookwala, Coppola, et. al. (2000) also document men's tendency to prefer life-sustaining treatments over palliative care, though they attribute it to men's and women's different experiences with the healthcare system over the course of a lifetime.

(Khader 2011, Khader 2009). In the second section, I argue that changing such preferences, though a generally appropriate goal of public policy, is generally not an appropriate goal for medical personnel involved in end of life decisions. More important at the end of life is respect for the coherentist autonomy of patients. In the third section, I argue that autonomy concerns do warrant intervening in lower-order IAPs that either a) seem inconsistent with an agent's value system or b) are based on misinformation. I draw on feminist theories of relational autonomy to show that such interventions by practitioners and family can be autonomy-enhancing rather than paternalistic.

Before I proceed, I briefly explain what I mean by the term "inappropriately adaptive preference." An IAP is a preference that is causally related to unjust social conditions and that an agent who was successfully and coherently pursuing her basic wellbeing would not express. By "causally related to social conditions," I mean something broader than what Jon Elster meant in his original work on adaptive preferences – something more consistent with what contemporary political philosophers mean when they discuss adaptive preferences. For Elster, preferences that are causally related to conditions must have formed "behind the back" of the person who has them, as in the case of the farmer who unconsciously starts preferring the city to the country now that industrialization has forced him to move (Elster 1987, 109-133). Contemporary theorists of adaptive preference describe preferences Elster would not describe as adaptive; Amartya Sen describes actively deciding to stop protesting one's tyrannical government because one cannot change it (Sen 2002, 634), and Martha Nussbaum describes the preference to live without decent sanitation because one has never known a life with it (Nussbaum 2001, 113). I believe a preference is causally related to injustice if it would be reversed on reflection and exposure to better information and conditions. Readers wanting a full justification of this definition of IAP should consult *Adaptive Preferences and Women's Empowerment.* However, it should become clear in a moment why my definition of IAP, with its perfectionist content, is more plausible than autonomy-based definitions.

6.2 Higher-Order Normative Adaptive Preferences as Compatible With Coherentist Autonomy

The paradigm cases of adaptive preference formation involve a causal relationship between the agent's circumstances and her values. Consider the classic example from La Fontaine's fable of the fox and the grapes. The fox wanted them (ie, thought they were valuable) until he he could not have them. Now he thinks they are sour but he is unaware of the reasons for his preference change. I am about to argue that the fox can make autonomous decisions on the basis of his newfound dislike for grapes. Before I do, it is useful to make a certain distinction – between autonomously f*orming* a preference and being able to autonomously make decisions *on the basis of* a preference. I do not dispute that the fox came to

devalue the grapes through nonautonomous processes. After all, he does not really know *why* his preferences changed; he may not even remember *that* they changed. I do happen to be skeptical of the idea that it would be possible or desirable to hold only autonomously *formed* preferences, [2] but we need not take a stance on that question to hold that the fox's preference, though nonautonomously formed, can form the basis for other autonomous preferences.

If the dislike for grapes becomes a part of the fox's deep values, it is fully possible for him to make future procedurally autonomous choices on the basis of it. I want to offer a richer example – partly so that it is easier to relate our discussion to end of life decisions – and partly because of the contrived character of discussing the deep values of a fox. So let us turn to a relevantly similar example concerning elderly woman with a terminal disease. I will call her "Lucy." Lucy's mother was a full-time caregiver and so was she. Lucy was raised to believe that women and men had different roles, and that women's moral duty was to take care of others. Her parents and gender ideology prevented her from training for a career outside of the home, and her husband would not have permitted her to work outside the home if she wanted to. Lucy was told her purpose in life was to nurture others – to make sure her husband was fed and listened to and had a clean house in the evenings, to make sure her children and grandchildren were safe and loved. As an adult, Lucy believes she has chosen this form of life and that the values it entails are the correct ones. She has even felt happy living her life this way. Now, at the end of her life, Lucy is incapacitated such that she cannot care for anyone. She believes her life can have no meaning now and that it is thus time for her to accept palliative care and die. I take Lucy's preference structure to be relatively common among – but not necessarily paradigmatic for – those who have internalized oppressive beliefs.

Lucy's case has two key things in common with the fox's. First, Lucy's adaptive preference has a normative dimension; it concerns her values. Let us also suppose that it is like the fox's in that it affects her values at a high order – though it is unclear what order of preference grape-liking is in foxes (readers who think the fox has adapted only a lower-order preference will find the case of Patty below more analogous). Second, Lucy's adaptive preference, like the fox's, was initially formed without her awareness or choice. She did not ask to have a mother who was a full-time caregiver, to have parents who endorsed patriarchal values so thoroughly, or to live in a patriarchal society – and by the time she could reason about this lack of choice, some of her core commitments had already been formed.

[2] Briefly, my reason for skepticism of the idea that all of our preferences should be autonomously formed is that it seems incompatible with certain facts about personal development. All of us form preferences before our deliberative capacities are well-developed enough to scrutinize them. We can acknowledge this without denying that it is desirable to scrutinize and autonomously endorse or reject such unreflective preferences *after* they have formed.

Elster wants to claim that the unconscious nature of such choices renders them nonautonomous but says strangely little about the theory of autonomy on which this could be the case. It cannot be mere unawareness of a preference's formation that renders a preference nonautonomous; I did not consciously decide to avoid physical pain, but we do not say that this preference is nonautonomous.

There are two plausible ways of construing Elster's underlying claim about autonomy – neither of which generates the conclusion that Lucy's preference for palliative care is nonautonomous in the sense relevant to end of life decisions. According to the first, Elster's underlying claim about autonomy is restricted to cases of preference *change.*[3] Elster may be agnostic on the question of how autonomous preferences are initially formed. His point may simply that, once preferences have been formed, changes to them that occur "behind the agent's back" are nonautonomous. Such a view, however, cannot yield the conclusion that Lucy's preference for palliative care is nonautonomous. Lucy did not once prefer life-sustaining treatment and experience an unconscious reversal of that preference. Rather, Lucy consulted her values to form the preference – values that she has reflectively endorsed.

Ben Colburn suggests a second way of understanding Elster's underlying claim about autonomy. Colburn (2011, 65-67) usefully distinguishes between preferences whose genesis is *necessarily* unconscious and whose genesis *happens to be* unconscious. Preferences in the former category originate in what he calls "covert influence." The autonomy-related problem with such preferences is that they cannot be a part of attempts to decide for oneself what is valuable (Colburn 2011, 62). According to Colburn, the fox's dislike of grapes is nonautonomous because of the necessary contradiction between the fox's reasons for believing he dislikes grapes and the true reason. The fox believes the grapes are sour, but the fox could not coherently hold this belief if he knew he had falsely attributed sourness to the grapes to protect himself from dissatisfaction. Some preferences influenced by unjust social arrangements are also necessarily nonautonomous, and Lucy's higher-order preferences may fit the bill.

It is, of course, much harder to argue that Lucy's values are mistaken than it is to argue that the fox is mistaken about the taste of the grapes. Let us suppose for the sake of argument that Lucy holds the values she holds because of limits on her opportunities, despite her belief she holds them because of their objective truth. It is self-contradictory for Lucy to believe that she holds her values because they are true and to be aware that she holds them because of contingencies related to how she was raised. If Lucy's relation to her sexist values is necessarily nonautonomous, the decions she makes on the basis of those values – such as the decision to pursue palliative care – are also nonautonomous. Clearly, not all sexist values acquired in sexist societies will have the necessarily nonautonomous character Colburn describes. For instance, a woman may, more or less coherently, endorse sexist values simply *because* she was raised to believe in them. Still, it

[3] I thank Juha Räikkä for suggesting this interpretation of Elster to me.

seems correct to me to argue that many real-world women have internalized sexist beliefs through the covert influence of ideology and restrictive options.[4]

We might ask whether Colburn's understanding of preference autonomy is generally plausible. Though I have my suspicions that it is not, I will concede for the sake of argument that many preferences adapted to injustice are formed through processes that Colburn (and if we accept Colburn's interpretation, Elster) would deem nonautonomous. It does not follow from the view that Lucy's higher-order preferences are nonautonomous in Colburn's sense that Lucy's preference for palliative care should be questioned or overridden. Colburn's understanding of preference autonomy is not the conception relevant to gauging the autonomy of end of life decisions. The type of autonomy we should strive to respect in end of life decisions should have presumptive authority is what we might call "coherentist autonomy."

I will offer reasons for this focus on coherentist autonomy in the next section. Now, I want to show that decisions like Lucy's to enter palliative care are autonomous on coherentist conceptions. Coherentist conceptions, also known as "endorsement," "identification," or "integrative" conceptions are a subset of procedural conceptions of autonomy. As such, they state that preferences may be autonomous irrespective of their content. I may choose autonomously to seek life-sustaining or palliative treatment; what makes the choice autonomous is how I stand in relation to my decision. Broadly understood, coherentist theories hold that my actions are autonomous when they are reflective and in harmony with my other network of values and desires. Theorists I would place in the coherentist camp include Michael Bratman (2007) Gerald Dworkin (1988), Harry Frankfurt (1988), and Marilyn Friedman (2002). Exponents of coherentist understanding of autonomy differ about what constitutes the perspective with which my autonomous choices must harmonize.

Coherentist conceptions of autonomy have no problem with choices like Lucy's treatment decision. Lucy's belief that her purpose in life is to give care is part of her higher-order value set, and this motivates her not to prolong her life. We may attempt to show otherwise – that Lucy's palliative care choice is nonautonmous – by referring to coherentism's emphasis on endorsement. That is, we may argue that the preference to fulfill an oppressive gender role cannot be a part of an autonomous self-conception if it was not endorsed during its formation. Lucy did not reflect on and endorse repressive gender norms while she was beginning to

[4] One piece of evidence for the idea that sexist socialization can occur through covert influence could be women's descriptions of the experience of consciousness-raising. Feminists often recall having abandoning sexist values upon realizing that their true genesis was something like indoctrination. However, an argument that consciousness-raising works because it reveals the true source of sexist beliefs relies on the metaphysical assumptions that beliefs have single, identifiable sources and that women's narratives correctly identify their sources. There are certainly other ways of understanding what happens in consciousness-raising–for instance, as involving changes in women's values rather than mere exposure of the conditions of their genesis.

believe in them. However, this argument misunderstands what is required for endorsement. Most theories of autonomy hold that endorsement requires only reflecting on and accepting a view – and this may occur *after* one has formed the view. Since the adult Lucy has thought about and accepted her belief in the rightness of oppressive gender roles, that belief passes coherentism's test for factoring into autonomous decisionmaking.

We might also attempt to demonstrate that Lucy lacks coherentist autonomy by arguing that, though Lucy is ruled by higher-order preferences, her higher-order preferences are not true ones because they contradict her deep values. Marilyn Friedman (1986) argues that oppression cases are precisely the ones that show us that being ruled by *higher-order* preferences and having *harmonious* preferences are not the same thing. Friedman's example involves a woman who has the higher-order preference to be a dutiful wife but lower-order feelings of depression, anxiety, etc. regarding this role. Friedman suggests that the more autonomous choice may be to act on the lower-order frustrations and reform higher-order desires.

I would say that a coherentist theory of autonomy can support the intuition that rejecting higher-order preferences is more autonomous in exactly two sorts of cases – neither of which is Lucy's. A person can be autonomous and reject a higher-order desire if that higher-order desire is out-of-sync with a host of her other desires or her deeply felt sense of self. Say the woman in Friedman's example wants to be a dutiful wife but also wants not to do the dishes, hates housework, wants to earn an independent living etc. Perhaps the joy she feels in not doing the dishes reveals a set of previously unacknowledged and inarticulate desires inconsistent with housewifery. Or, a person can autonomously reject her higher-order desires if she happens to have a desire of a particularly high order that she has previously failed to integrate into her self-conception. In this second sort of case, the housewife would have to have a desire for self-actualization that was of a higher-order than the desire to be a "good woman".[5] But, though many women socialized into caregiving probably have desires inconsistent with their lower-order expressed values, Lucy's case is not like this. Her higher-order belief that caregiving is her purpose is consistent with a wide variety of her beliefs and behaviors.

Another way to attempt to show that Lucy's treatment decision is objectionable on coherentist grounds would involve invoking criteria for procedural independence. Most coherentist theorists argue that autonomous preferences are not only consistent with the agent's values, etc. but also arrived at reflectively and absent coercion and manipulation from other persons. Now, I have conceded for the sake of argument that we can think of Lucy's higher-order preferences as having been problematically arrived at. However, I think we should be wary of thinking of Lucy's preferences as having been *coercively* arrived at. Such a view would yield the strange conclusion that all preferences generated through

[5] This is, of course, not the explanation Friedman herself favors, since her point in the examples is to undermine hierarchical accounts of autonomy.

socialization and later reflection are coercively arrived at. More important for our current discussion, however, is that Lucy's decision to pursue palliative care was made free from coercion or manipulation by others. Still, it may be argued that lack of self-worth affects procedural independence – even if it is not coercively instilled. I believe this involves importing substantive content into autonomy in a way that is ill-suited to end of life decisionmaking. Readers interested why we should not focus on substantive autonomy at the end of life should consult the sections of the paper on MacKenzie's and Oshana's conceptions of autonomy in the next section of the paper.

Let us examine a final coherentist way we may attempt to show that Lucy's desire to forego life-sustaining treatment is nonautonomous. We may argue that Lucy would have repudiated her belief that caregiving was her sole role while she was forming that belief, if she had been aware of it. This argument presupposes that there was a time when Lucy wanted something more than this. At that time, her belief about women's subservience was inconsistent with other beliefs and desires she had. I do not find this suggestion implausible; I have argued elsewhere, that most people desire their basic wellbeing (Khader 2011). Perhaps, when Lucy was younger, she had a sense that she was valuable regardless of whether she could provide service to others. So perhaps *that* Lucy, the Lucy of many many years ago, would have been upset that she was forming preferences that made her self-worth so contingent. But the important point is that this is not what happened. What matters for coherence autonomy is not what would have cohered with the desires of a past version of an agent. What matters is what coheres with the desires of the present – or at least, recent – agent. Of course, there will be cases in end of life decisionmaking where a person's consciousness is fading and surrogates must refer to her past decisions to make decisions for her. But this is because she is no longer completely an agent. If we are still dealing with an agent, we care what actions the agent before us would repudiate. If Lucy's views about women's roles are now part of the core of who she is, it is coherent for her to want to end her life when she cannot continue caregiving.

Feminist theorists, and a wide variety of autonomy theorists with political concerns, think it is a defect of coherence conceptions of autonomy that they cannot find problems with preferences like Lucy's.[6] But I want to be clear, contra the worries of feminist critics of coherentism, that accepting coherentism does not require taking a sanguine view of Lucy's oppression. Lucy's preferences were probably harmful to her during her lifetime. She remained in a relationship with a person who did not view her as an equal. She missed opportunities to cultivate her other capacities; perhaps she might also have been a chemist or musician or athlete. She earned no income during her life, and this put her at risk for poverty, homelessness, and a variety of other ills. Her focus on serving others may have caused her to compromise her physical health; indeed, there is plenty of evidence that informal caregivers are at risk for poor health outcomes (Bevans and

[6] See Stoljar (2001).

Sternberg 2012). Further, her inability to locate her value in anything but serving others may have actually damaged her ability to experience connection with those who love her. In short, Lucy's internalized oppressive beliefs probably produced a number of limitations on her capacity to lead a good life.

But the tragedy in Lucy's case can be articulated using concepts besides coherentist autonomy. I share a common "feminist intuition" (to use a term from Stoljar) that something is wrong when unjust social institutions encourage oppressed people to develop self-harming behaviors. But we need not argue that there are autonomy harms to show *something* is wrong. Preferences like Lucy's attack the perfectionist value of living out one's potential, the values associated with care and reciprocity, and the social value of justice. There is no doubt that social institutions ought to have offered her a wider range of options and helped her cultivate a less precarious sense of self-worth. I also believe that the adult Lucy was owed opportunities to expand her sense of what she was capable of. But the question with regard to end of life decisions is not about how to improve a person's long-term future. As I now turn to arguing, perfectionist values, and values of justice and reciprocity are not the values medical professionals should attempt to promote in end of life decision making.

6.3 Why Focus on Coherentist Autonomy at the End of Life?

There is relatively wide agreement that medical professionals involved in end of life decisionmaking ought to focus on promoting patient autonomy. Even critics of the focus on autonomy in the mainstream bioethical literature tend to argue that something is wrong with our *operative conception* of autonomy – rather than that autonomy is an inappropriate value.[7] I have argued elsewhere, however, that liberal public institutions can be justified in intervening to change socially-induced preferences with harmful content, regardless of whether those preferences are autonomously formed or endorsed. My reasoning for this claim has been perfectionist – that one appropriate aim of the state is to promote people's basic wellbeing. In line with other liberal perfectionists, I have argued that state action on perfectionist values is acceptable when three conditions are met: a) the perfectionist values are not aligned with any particular comprehensive doctrine, b) the intervention is designed to promote only basic levels of welfare, and c) the intervention is noncoercive. If these claims about how liberal societies can

[7] For instance, Winzelberg, Hanson, and Tulsky (2005) argue that medical professionals overvalue autonomy in end of life decisions. However, their argument against the contemporary focus on autonomy seems to be that patients do not want doctors to focus on autonomy. According to their reasoning, doctors should try to help patients "receive care consistent with their preferences" (2005, 1046). It is difficult to see why this would be a value without some underlying value for self-determination. What Winzelberg, et. al. really seem to be arguing is that medical ethics focuses excessively on procuring consent for each individual treatment decision.

promote perfectionist values are true, we need an argument for why autonomy should prevail over perfectionist values in end of life decisions – one that is compatible with liberal perfectionist assumptions. We need to know why medical professionals should not challenge preferences like Lucy's – given that the values on the basis of which Lucy chooses palliative care are values harmful to her. Before I proceed to defend a focus on coherentist autonomy at the end of life, I note the limitations of recommending such a focus. Clearly, patients whose age or diseases are advanced such that they lack core agentic capacities cannot have their autonomy promoted, per se. In these cases, I would suggest that attempting to act in accordance with a patient's values are still of high importance.

I see three reasons medical professionals should avoid focusing on perfectionist goals at the end of life. Examples of such perfectionist goals would include trying to undo the force of sexist or racist ideology, trying to prolong (or end) life, or trying to get the patient to take up a particular (purportedly correct) perspective on her purpose in life, etc. First, promoting such goals, when it involves asking people to change their higher-order preferences (as it would in Lucy's case) is likely to produce welfare losses that will not be outweighed by future welfare gains. IAP interventions are justified during the course of life on the grounds that they allow people to actualize potential that unjust social arrangements have stymied. We hope the potential for future welfare outweighs the utility cost involved in scrutinizing – and often abandoning – some existing preferences. Lucy would probably have experienced losses at being asked to change her views about her purpose in her 20s; conflict may have arisen in her relationship, she may have felt confused and unmoored for a while, and so forth. The likelihood of such costs being overridden by future benefits, though significant when a person faces a long future, diminishes at the end of life. This is both because of the shortness of the time frame and the level of attachment to her higher-order preferences (inappropriately adaptive or otherwise) a person has likely developed by the end of life.[8] Asking Lucy to change her views in her 90s will probably involve more than a temporary sense of unmooring. It is likely to result simple exhaustion on Lucy's part (and Lucy is probably already experiencing pain and exhaustion) or in a deep loss of sense of self.

This argument against medical professionals attempting to change preferences adapted to injustice at the end of life is fully compatible with perfectionist views about welfare. The idea is that such attempts are likely to cause decreases in welfare. It may be objected that the definition of welfare operative in my argument is hedonist and thus one that perfectionists should reject. To say that it will cause suffering to uproot a person's existing values would seem to value subjective welfare over objective welfare. It may seem that perfectionists are bound to say that Lucy is better off if she stops endorsing sexist beliefs – regardless of whether this produces grief, sadness or feelings of loss on Lucy's part. I do not believe

[8] The part of the argument about deep attachment to one's preferences being common in end of life decisions applies most readily to patients who are elderly as well as terminally ill.

perfectionism must yield this conclusion. Subjective wellbeing may be one part of flourishing, and, if this is the case, it is possible for an agent to suffer so much subjective pain that it outweighs the value of the functionings she gains. This, I would submit, is what is likely to happen to Lucy.

Second, questions about what to do at the end of life are metaphysically controversial and do not admit of simple answers. That is, it is not obvious in most cases whether it would be better for a person to end her life, or pursue life-prolonging or palliative care. To say that end of life decisions are genuinely controversial is not to deny that many agents making them are motivated by genuine IAPs. Lucy's belief that women's purpose is to serve others is clearly incorrect, harmful, and one that a liberal society ought to have discouraged. But what Lucy should do at the end of life is far less clear – raising, as it does, questions about the ultimate meaning of human existence, Lucy's beliefs about an afterlife, and so forth. Given this, there is no clear moral obligation on the part of medical professionals to promote a particular choice. If we are not sure what the good is, we certainly cannot claim that medical professionals have an obligation to promote it. Our certainty that higher-order preferences like Lucy's are incorrect does not change the fact that it is unclear which lower-order decisions are the right ones. Given the welfare losses typically involved in trying to change such preferences, we would need a separate argument for attempting to change lower-order treatment decisions. Put in terms of our example, if we know that Lucy will suffer immensely from the unmooring of her sense of self, we need some reason to believe that changing her desire to die in comfort will ultimately be good, or good for her. Such a reason does not appear forthcoming.

This brings us to the third reason medical professionals ought not focus on perfectionist or justice-related goals at the end of life. Such a focus would be inconsistent with respect for diverse conceptions of the good. It is acceptable to promote perfectionist ends in a pluralist society when those ends are not wedded to any particular vision of the good life. But virtually any conclusive view about what someone in Lucy's position should do will draw on a comprehensive doctrine. For instance, the option of physician-assisted suicide will look particularly horrible to a person who believes that those who commit suicide experience eternal damnation. The same option will look desirable to a person who believes that the good is to "live life to the fullest." The comprehensive doctrines to which both of these views belong are common in Western societies. Medical professionals who promote a particular course of action at the end of life are likely to be foisting a comprehensive doctrine onto patients.

But why promote *coherentist* autonomy? We can begin to answer the question by noting that most bioethical literature emphasizes self-determination, where self-determination is understood as the ability to make decisions free from undue influence from others. Arguments for self-determination typically presume the value of coherentist autonomy. The clearest reason for valuing self-determination in medical decisions is that it allows the agent to make medical decisions in a way consistent with her values. We can see this in bioethicists' tendency to justify their recommendations on the grounds that patients will be able to choose treatment

plans in line with their desires (Ho 2007). We can also discern a value for coherentist autonomy at the end of life in discussions of the values that are typically recommended to guide surrogate decisionmaking; we expect surrogates to try to decide in a way consistent with the patient's values and beliefs.[9] Common claims about the conditions under which normatively-focused intervention by medical professionals is acceptable also presuppose the value of coherentist autonomy. For instance, it is commonly argued that medical professionals may dig deeper when patients express preferences contrary to their values, because inconsistent preferences may reveal impediments to patient self-determination (see Brudney 2009, 31-32, Childress 1990, 14).

Certainly, these claims about the existing focus on self-determination and coherentist autonomy at the end of life do not amount to a philosophical defense of those values. We can get closer to a defense by discussing the reasons we should not promote the two key competing conceptions of autonomy at the end of life. One competing sort of autonomy is the one advocated by Colburn above; call this "self-authorship" autonomy. If we value autonomy as self-authorship, we should look to see whether an agent's decision and the values motivating it have been acquired through processes she could, in principle, be aware of. It is worth bearing in mind that, on this conception of autonomy, Lucy's values were nonautonomously acquired. She thinks she values caregiving so highly simply because of its objective virtues, but if she were aware of her own socialization, she would have to see that limited options have also influenced her values. Two important considerations weigh against promoting this sort of autonomy in end of life decisions.

The first is practical. Both the nature of socialization and the parameters of end-of-life medical encounters make it difficult to ascertain whether people's values have been influenced through processes that could not in principle be endorsed. Even if oppressive socialization bears all the trademark features of covert influence, its presence or absence is much more difficult to identify than those of the more paradigmatic examples of covert influence such as brainwashing, subliminal advertising, etc. This is partly because oppressive socialization, unlike subliminal messaging, does not occur as part of a discrete event and partly because it frequently occurs at ages when people's deliberative capacities are not in full effect. The end of life medical context compounds the difficulties that would be involved in discerning the true reasons a person acquired her oppressive values.

[9] Daniel Brudney argues that the value of authenticity should guide end of life decisionmaking and that this is a distinct value from autonomy. But, given that authenticity involves acting harmoniously with one's values, it seems identical with coherentist autonomy–at least when sought by the agent herself. The difference between coherentist autonomy and authenticity seems to arise in the case of surrogate decisionmaking–where a third-party tries to make decisions on the basis of the patient's values. The purpose of recommending that surrogates act in accordance with a patient's values seems to be continue to respect the patient's right to self-determination after she has lost agentic capacities.

The patient probably acquired her values very long ago, and there is not an easy way to obtain sufficient information about the conditions of initial acquisition.

A second reason to reject self-authorship autonomy at the end of life concerns a value for respecting patients' agency. Medical professionals who are convinced that patients are under the influence of oppressive socialization, covertly delivered, may eviscerate patients' senses of self. I noted above that there are perfectionist reasons to worry about the unmooring that may result from attempts to change people's higher-order preferences at the end of life. I also believe such unmooring can be construed as problematic on any conception of autonomy. Virtually any definition of autonomy will concede that a person cannot act autonomously unless she really cares about something. Autonomy theorists are quick to point out the difference between the autonomous person and the one who is simply buffeted by fate. The danger of telling a person that she is wrong about who she is that she may lose her sense that she is an integrated self. This loss of self registers as a harm to autonomy as well as a harm to welfare.

The self-authorship conception of autonomy is not the only one that would deliver the conclusion that agents like Lucy have nonautonomous preferences. A variety of substantive conceptions of autonomy would also support the conclusion, so we should also consider them as competitors to coherentism. Substantive conceptions of autonomy stipulate that choices cannot be autonomous if they do not express certain moral commitments. Marina Oshana advocates a substantive theory that is heavily content-laden – and that is explicitly designed to recognize the autonomy harms of oppressive socialization. According to Oshana (2003, 251), the preference to live one's life consistently subject to the will of others, regardless of how it is formed or viewed by the agent, cannot be autonomous. Oshana's example is of an Afghan women who endorses the Taliban's various restrictions on her mobility, economic opportunities, and self-expression. Suppose our Lucy's preference for repeated self-sacrifice is similar; her day-to-day actions have been mostly dictated by others because she has located her entire self-concept in serving them.

Despite Oshana's recognition of the force of social injustice on individual preferences, it is unclear that her conception of autonomy can be justifiably promoted at the end of life. Unsurprisingly, given its perfectionist content, it is objectionable on the same grounds that other perfectionist interventions in end of life decisions are. Oshana's is a theory of *personal* autonomy; it is clearly a view about what produces an autonomous *life*. Oshana seems to believe that a life lived without repeated self-determined decisions is one that fails to actualize important human potential. The goal of promoting strong substantive autonomy seems to derive from the value of living a certain sort of life – but it is a form of life that will likely never come about for the person at the end of a terminal illness. It is far from clear that the values that enable one to flourish over a lifetime are the ones that matter at the end of life. Indeed, the end of the struggle with a terminal illness often involves coming to terms with increasing biological limitations on one's decisionmaking competence; asking a person to stop submitting to the will of others at the moment when she is losing her ability to form and execute her own

may even be misguided. Second, the strong substantive conception of autonomy is inconsistent with diverse conceptions of the good. Interestingly, Oshana (2003, 106) herself acknowledges this – arguing that it is wrong to expect all conceptions of the good to be compatible with autonomy. Oshana's position that only some forms of life can be autonomous is certainly plausible. But, with respect to our current set of questions, it amounts to a concession that medical professionals who promoted strong substantive autonomy at the end of life would be foisting a somewhat comprehensive doctrine on patients.

A weaker substantive conception of autonomy would be less subject to these criticisms. Many feminist theorists hold weaker substantive conceptions of autonomy, on which choices must express self-regard to be autonomous.[10] Mackenzie, whose view I will take as the exemplar here, argues that an agent cannot successfully reflect on who she wants to be and what she wants to do without believing her values and actions are worthwhile (MacKenzie 2008, 525). The problem with promoting this type of autonomy at the end of life is that, depending on how we interpret it, it may foreclose consideration of the very question that the patient at the end of a terminally illness is trying to answer. MacKenzie's understanding of autonomy does not stipulate whether the agent must *subjectively feel* that her life is worthwhile – or whether she must believe that her life has a meaning. The former requirement may make sense in the case of a functioning person; we believe that the functioning person is wrong when she says her life is worthless – and usually see such claims as a symptom of depression. But is impairment the only reason the terminally ill patient may believe her life is – or will soon be – no longer worth living? If the autonomous agent must feel her life is worth living, the decision to die may never be autonomous.

Alternatively, MacKenzie's self-worth condition for autonomy may mean that the agent must believe her life has a meaning. This would not categorically rule out the choice to die; dying may even be the way of giving my life a meaning. I have no objection to medical professionals promoting weak substantive autonomy, construed in this way. But here, the substantive content is weak enough as to already be implicit in coherentist conceptions. Coherentist conceptions demand reflective integration – not just some sort of accidental consistency among the agent's values and behaviors. The difference between attempting to attribute meaning to one's life and acting on the basis of a sense of who one is begins to seem semantic. Once we recognize the similarity to coherentist views, we see that a weak substantive conception of autonomy would not justify attempting to change higher-order IAPs anyway. For a sense that her life has a purpose is something Lucy actually has; the purpose has just already been fulfilled through service to others.

6.4 When Coherentist Autonomy Justifies Intervention

Coherentist autonomy concerns are appropriate at the end of life and cannot justify attempts to change preferences like Lucy's. But we should remember that not all

[10] See MacKenzie (2008), Superson (2005), MacLeod (2002).

cases in which IAPs motivate end of life decisions are like Lucy's. Lucy's IAPs are normative, higher-order, and relatively thoroughgoing. Unjust social conditions do not only affect people's preferences by inducing pervasive or higher-order value distortion. They may induce mid- or low-level value distortion – or no value distortion at all.[11] Where adaptive preferences of these orders affect end of life decisions, there may be coherentist autonomy-based grounds for intervening.

Consider the case of Patty, whose medical interactions are shaped by a different type of gendered IAP. When Patty was growing up, those around her did not place much value on women's formal education. She was taught to be deferential to people in positions of authority – almost all of whom were men. But as she grew into adulthood, she began to believe that she was smart and competent in ways that people did not recognize. Her belief in her own intelligence is a direct result of her competent responses to the adversities she faced – responsibilities that surprised her. She sees herself as a "fighter" and believes that the most important thing in life is never to give up. She has raised her daughters to believe that they, and other women, are competent to be entrusted with responsibilities. However, she struggles to behave in a way consistent with her beliefs. Her early conditioning still shows itself in her deference to male authority figures. In speaking about her interactions with them, she will occasionally say things like, "What do I know? I am just a woman." Now, toward the end of a terminal illness, Patty's options are relatively short-term palliative care and six months of life-prolonging care that may involve slow and painful degeneration. Patty believes her doctor has recommended the palliative option and appears resigned to complying.

Patty, like Lucy, has internalized values that are harmful to her as the result of sexist socialization. But, unlike Lucy's, her highest order values are not ones that exposed her to harm during her life. She has a pesky value for deference to male authorities, but it is inconsistent with the rest of her values. Because of this, intervention in Patty's case need not be justified on the basis of perfectionist concerns. If we understand autonomy in a coherentist fashion, we have good reason to suspect that Patty's readiness to consent to palliative care is not autonomous. Her higher-order beliefs that women are competent, and her higher-order value for "fighting" seem wildly inconsistent with the course of action toward which she is leaning. To be clear, suspecting Patty's preference of heteronomy entails no moral judgment about the value of palliative care. Refusing to fight may be the morally superior option, but the point is that it is not the one that *Patty* is disposed to see as such. The strength of any medical professional's warrant to ask Patty to further scrutinize her decision will depend on the extent of her acquaintance with Patty and her values. But, where medical staff have reason to believe Patty is acting inconsistently with her values, there is reason to question Patty's choices.

[11] I have attempted to develop a conceptual repertoire for describing the variety of ways unjust social institutions may affect the preferences of the deprived in Khader (2012).

Consider another case in which a person's preferences have adapted to injustice without producing higher-order value distortion. Linus, also an elderly person at the end of a terminal illness who is deeply afraid of being a burden to his family members at the end of his life. He spent his life being the male breadwinner in a "traditional" nuclear family. He and his wife have five children, whom he loves. He does not hold particularly sexist values, but he spent only occasional time with his children when they were young. His wife and children are more than willing to take care of him, but he is having trouble grasping this fact. He imagines that dependent others are only burdens to those who care for them, and this is largely because he has little experience as a caregiver. He values his own wellbeing but believes that, at the end of life, he should choose in the way that will minimize suffering for his family. He is seriously considering opting for physician-assisted suicide to avoid being perceived as worthless by his family. However, his family is fully willing to care for him and would value the experience of spending extra time with him.

It may be unclear to some why Linus' apparent preference is causally related to unjust social arrangements. Linus struggles to grasp the idea that dependent people are not pure burdens for a particular reason – because of having occupied a traditional male gender role in his society. The idea that the experience of caregiving shapes people's beliefs and values offers a plausible explanation of the (perhaps surprising) empirical data suggesting that men worry more about the experience of being dependent on others. A number of feminist theorists argue that women's experiences as caregivers make them more able to appreciate the role of dependency in the human condition.[12] Many feminist theorists also argue that the gender division of labor, and particularly the invisibility of women's dependency work, encourage men to buy into the idealization that human beings are self-sufficient.[13] To be precise about the relationship between Linus' apparent preference and social injustice, Linus' degradation of dependency has not been harmful *to him* in his life – though it may have harmed the women in his life. He does not have what I would call an IAP, but he makes his end of life decision on the basis of preferences influenced by an unjust gender system.

I believe that Linus' case is another one where there are autonomy-based grounds for intervention. He is primarily wrong about a matter of fact – a fact that gender ideology has made it difficult for him to see. This is the fact that his family members will not perceive him only as a burden. This non-normative error causes him to be mistaken about the course of action that most closely aligns with his values. Linus's autonomy would be enhanced by an intervention to increase his awareness of how his family really views him. Gender ideology clouds Linus's ability to determine which course of action would best cohere with his values – and is thus autonomy-impeding in the coherentist sense.

In cases where unjust social conditions cause a person to be wrong about important matters of fact (Linus's) or hold selectively distorted lower-order values

[12] See Ruddick (1995).
[13] See Bartky (1991), Kittay (1999).

(Patty's) coherentist autonomy weighs against treating initial treatment preferences as autonomous. However, it is worth mentioning that my argument for intervening in treatment preferences like Linus's and Sally's authorizes a relatively wide scope of interventions in patients' initial treatment preferences. Coherentist autonomy in principle at least justifies questioning all initial inclinations about treatment that seem inconsistent with the patient's higher-order preferences – regardless of whether they have been influenced by social injustice. Those who want to justify a specific focus on interrogating lower-order preferences caused by unjust social arrangments will not find any principled reason for such a focus within the conception of coherentist autonomy. One way to achieve such a focus would be to supplement coherentist autonomy with a perfectionist view of what human nature causes people to want at the end of life. We could, for example, argue that people really want to stay alive at all costs, or to end suffering, and then argue that any treatment preference inconsistent with this is likely to be nonautonomous. However, this approach would be subject to all of my previous criticisms of perfectionism in end-of-life decisions. It bears repeating that any judgment like "people naturally want to stay alive at all costs" is too controversial to form the basis for any sort of policy.

There may, however, be some prudential reasons for focusing on apparently heteronomous treatment preferences related that seem adapted to social injustice. For instance, treatment choices based on preferences adapted to injustice may be particularly easy to spot. One reason for this could be is that preferences adapted to injustice tend to strongly track social group membership. For example, men seem particularly likely to be worried about being a burden. If data consistently revealed this phenomenon, and revealed that it was underlyingly motivated only by men's false beliefs about what their families would do for them, medical practitioners would have special reason to worry that men who wanted to end their lives because such worries were acting nonautonomously. Assuming other sorts of higher-order preferences do not accrue to patients in such a socially patterned fashion, and that medical professionals often lack intimate knowledge of each of their patients, it is more likely that medical practitioners will be right in suspecting the nonautonomy of gender-related treatment preferences. Still those of us who are particularly concerned about gendered adaptive preferences may have to be satisfied with the fact that coherentist autonomy provides only a blunt instrument for dealing with them – and deals only with a certain class of them. Or, we may simply concede, in end of life situations at least, that nothing makes lower-level gendered adaptive preferences more problematic than other lower-level heteronomous preferences.

My claim that autonomy justifies intervention to ensure that people's apparent preferences are actually theirs will seem puzzling to some. It is frequently thought that nonintervention is the best way to respect patient autonomy. However, as a variety of feminist bioethicists have pointed out, nonintervention only respects patients who are already free of impediments to their autonomy. The operative assumption seems to be that medical professionals are the only ones who could impede patient autonomy. But our discussion has revealed ways in which social

influences outside the medical context can impede the autonomy of medical decisions. This means that promoting the autonomy of patient decisions may involve more than simply presenting options and standing back; it may involve actively discussing the ways certain options fit into patient's existing views – normative and otherwise. Of course, the extent to which medical practitioners are qualified to do this will vary; they may not know enough about some patients' existing values to reasonably suspect that they hold preferences that do not fit. But in cases where medical staff do know, helping a patient act consistently with her values may involve helping her get clearer about certain non-normative facts (such as questions about her family's perception of her burdensomeness) and asking her to reexamine the consistency of her apparent choices with her normative beliefs. Of course, at the end of life, such conversations must be handled sensitively – so as not to unmoor or cause undue suffering to the patient. But we should not completely rule out the possibility that medical professionals can play an autonomy-enhancing role, especially when a meaningful relationship with the patient predates the end of life decision.

The idea that patients may have their autonomy enhanced through conversations about the coherence of their treatment plans with the facts and their values also lends credence to another point made by feminist theorists of relational autonomy – that there is sometimes an important role for family in end of life decisionmaking. Those who are close to a patient are frequently – but not always – well-situated to help the patient clarify the relationship between a proposed treatment plan and the rest of a patient's beliefs. As Anita Ho argues, families may have a particularly important role to play in helping patients orient themselves in a medical context many often find intimidating and alienating (Ho 2008, 131). Consider the case of Patty above. Patty opts for palliative care largely because of a type of deference to male authority the medical context has triggered in her. A family member of Patty's may be particularly good at spotting this and help Patty to make a decision more consistent with her values. Similarly, Linus' family members may be the ones best situated to let him know that his understanding of the burden he will impose is distorted. This is not to say that the influence of family members is always autonomy-enhancing. Some family members may use the medical system to abuse patients, or they may simply prioritize their interests over the patients'. However, as Anne Donchin argues, this seems to be an argument for encouraging medical staff to become acquainted with family dynamics rather than an argument for excluding family members from discussions about end of life decisions (Donchin 2000, 190).

Thus, a concern with coherentist autonomy justifies limited intervention by medical professionals and family to reverse preferences that are causally related to unjust social conditions. However, the justification offered by coherentist autonomy for intervention in preferences adapted to social injustice is accidental. Autonomy concerns, on their own, do not dictate that we should be particularly concerned about uprooting preferences that are adapted to unjust conditions. Particular concern about preferences that are causally related to injustice is justified by perfectionist and justice concerns – concerns that aim at respecting

people's entitlements to a flourishing life. The pursuit of perfectionist concerns is appropriate in policies directed at shaping the long-term trajectories of people's lives. However, people's own values, regardless of their content, assume special importance with regard to end of life decisions.

References

Babbitt, S.: Feminism and Objective Interests: The Role of Transformation Experiences in Rational Deliberation. In: Alcoff, L., Potter, E. (eds.) Feminist Epistemologies, pp. 245–265. Routledge, New York (1993)

Bartky, S.: Feeding Egos and Tending Wounds: Deference and Disaffection in Women's Emotional Labor. In: Bartky, S. (ed.) Femininity and Domination: Studies in the Phenomenology of Oppression, pp. 99–119. Routledge, New York (1995)

Bevans, M.F., Sternberg, E.M.: Caregiving Burden, Stress, and Health Effects among Family Caregivers of Adult Cancer Patients. J. Am. Med. Assoc. 307, 398–403 (2012)

Bookwala, J., Coppola, K., et al.: Gender Differences in Older Adults' Preferences for Life-Sustaining Medical Treatments and End-of-Life Values. Death Stud. 25, 127–149 (2001)

Bratman, M.: Structures of Agency: Essays. Oxford University Press, Oxford (2007)

Brudney, D.: Choosing for Another: Autonomy and Best Interests. Hastings Cent Rep. 39, 31–37 (2009)

Colburn, B.: Autonomy and Adaptive Preferences. Utilitas 23, 52–71 (2011)

Crawley, L., Payne, R.: Palliative and End of Life Care in the African-American Community. J. Am. Med. Assoc. 284, 2518–2521 (2000)

Donchin, A.: Autonomy, Interdependence, and Assisted Suicide: Respecting Boundaries, Crossing Lines. Bioethics 14, 187–204 (2000)

Elster, J.: Sour Grapes: Studies in the Subversion of Rationality. Cambridge University Press, Cambridge (1987)

Frankfurt, H.: The Importance of What We Care About. Cambridge University Press, Cambridge (1988)

Friedman, M.: Autonomy, Gender, Politics. Oxford University Press, New York (2002)

Friedman, M.: Autonomy and the Split-Level Self. Southern J. Philos 24, 19–35 (1986)

Hayes, D.W., Rayner, L.L., et al.: Gender Differences in Prevalence of Depression Among Patients Receiving Palliative Care: The Role of Dependency. Palliative Med. 26, 696–702 (2012)

Ho, A.: The Individualist Model of Autonomy and the Challenge of Disability. Bioeth Inq. 5, 193–207 (2008)

Ho, A.: Relational Autonomy or Undue Pressure? Family's Role in Medical Decisionmaking. Scand J. Caring Sci. 22, 128–135 (2008)

Johnson, M., Lin, L., et al.: Patients' Perceptions of Physicians' Recommendations of Comfort Care Differ by Patient Age and Gender. J. Gen. Int. Med. 15, 248–255 (2000)

Khader, S.J.: Must Theorizing about Adaptive Preferences Deny Women's Agency. J. Appl. Philos 29, 302–317 (2012)

Khader, S.J.: Adaptive Preferences and Women's Empowerment. Oxford University Press, Oxford (2011)

Khader, S.J.: Adaptive Preferences and Procedural Autonomy. J. Hum. Dev. Capabil 10, 169–187 (2009)

Kittay, E.F.: Love's Labor. Routledge, New York (1999)
Mackenzie, C.: Relational Autonomy, Normative Authority, and Perfectionism. J. Soc. Philos 39, 512–533 (2008)
McLeod, C.: Self-Trust and Reproductive Autonomy. MIT Press, Cambridge (2002)
Nussbaum, M.: Women and Human Development: The Capabilities Approach. Cambridge University Press, Cambridge (2001)
Oshana, M.: How Much Should We Value Autonomy? Soc. Philos Pol. 20, 99–126 (2003)
Rietjens, J.A.C., Deschepper, R., et al.: Medical End-of-Life Decisions: Does its Use Differ in Vulnerable Patient Groups? A Systematic Review and Meta-analysis. Soc. Sci. Med. 74, 1282–1287 (2012)
Ruddick, S.: Maternal Thinking. Beacon Press, Boston (1995)
Sen, A.: Gender and Cooperative Conflicts. In: Tinker, I. (ed.) Persistent Inequalities: Women and World Development, pp. 123–149. Oxford University Press, Oxford (1990)
Solomon, L.M., Noll, R.C.: Physician-Assisted Suicide and Euthanasia: Disproportionate Prevalence of Women among Dr. Kevorkian's Patients. Gender Med. 5, 110–114 (2008)
Superson, A.: Deformed Desires and Informed Desire Tests. Hypatia 20, 109–126 (2005)
Winzelberg, G.S., Hanson, L., Tulsky, J.A.: Beyond Autonomy: Diversifying End of Life Decision-Making Approaches to Serve Patients and Families. J. Am. Geriatr Soc. 53, 1046–1050 (2005)

Chapter 7
Sour Clinical Trials: Autonomy and Adaptive Preferences in Experimental Medicine

James Rocha

7.1 Introduction

Whenever possible, life and death decisions should be made autonomously. A decision is made autonomously when an agent calmly reflects over her preferences, which she had some role in developing, and determines that her decision is best for her, given her circumstances. When a medical patient makes a life and death decision autonomously, it is the best decision for her circumstances, as determined by her. If a medical professional pushes a patient to make a life and death decision non-autonomously, it would be severely morally problematic since the patient would not have the chance to make the decision she felt was best, which should be her right since her life is on the line.

Though we can certainly agree on the value of autonomous life and death decisions, it isn't always easy to ensure such decisions are made autonomously. The proximity of death could alter an agent's deliberation and diminish the likelihood of an autonomous decision. Death is likely to alter an agent's set of feasible options by eliminating some options and making others look more attractive. This alteration could lead to the creation of what Jon Elster refers to as "adaptive preferences," which amount to subconscious alterations of the agent's preferences based on a change in circumstances, or the agent's knowledge about the circumstances (Elster 1983, 109-140). When an option that the agent previously preferred turns out to be unavailable, something changes in the agent,

James Rocha
Department of Philosophy and Religious Studies
Louisiana State University
Baton Rouge, LA
The United States
e-mail: rocha@lsu.edu

J. Räikkä and J. Varelius (Eds.): *Adaptation and Autonomy*, SAPERE 10, pp. 101–115.
DOI: 10.1007/978-3-642-38376-2_7 © Springer-Verlag Berlin Heidelberg 2013

usually without her realizing it, so that she no longer prefers that option or comes to prefer an available option much more strongly.

The adaptive preference change is both fortunate (since the agent's original preference would have been frustrated anyway) and troubling (since the agent's preferences are being altered by subconscious neurological processes that are not in her control). Due to the latter, Elster argues this agent cannot be autonomous (Elster 1983, 20-25, 128-140). We could imagine an agent who has a benevolent angel following him around, altering his preferences so that he can trivially find happiness in whatever he's about to do. Such an agent, given the angel's control over his preferences, could not count as autonomous. Similarly, the agent who acts for her adaptive preferences satisfies preferences that were not developed by herself through any intentional or even conscious process. Control from within undermines autonomy just as external control would.

Since adaptive preferences are inconsistent with autonomy, it is important to seek them out in contexts where autonomy is necessary or relied upon, such as the medical field. As we have seen, life and death decisions may lead to adaptive preferences, but ideally should be made autonomously. Thus, adaptive preferences could raise a significant worry in medical contexts. To focus this worry, I will examine how adaptive preferences affect autonomy in clinical trials for terminal diseases. Such cases highlight the issues that arise when adaptive preferences undermine autonomy in medical contexts. So, while I shall focus on clinical trials, the conclusions will hopefully provide insight into a broader array of scenarios.

The scenario we will examine will involve an agent's preferences adapting to add value to an experimental drug that the doctor doesn't think she is likely to choose autonomously. I will argue that there are multiple ways in which a doctor may disrespect the patient's autonomy in this scenario. If the doctor withholds information, that could disrespect the patient's autonomy since the patient would be making a decision that wouldn't be adequately informed. If the doctor provides the information about the drug, she may be tempting her patient to choose a drug that he would refuse were he autonomous. Though these problems seem to leave the doctor no permissible option, I will argue that there are contexts where each is the correct choice, and thus doctors can manage to avoid disrespecting their patients' autonomy in these difficult adaptive preference scenarios.

7.2 Terminal Clinical Trials and Adaptive Preferences

Consider the following scenario. A doctor must choose between two drugs for a patient with a terminal disease. Drug Safe has been approved for use, will lessen the patient's suffering, will prolong his life for a short period, but it will not cure the disease and the patient will soon die. The doctor has also been asked to submit patients for a clinical trial for Drug Risky. In the first phase of the trial for Drug Risky, a non-negligible number of patients had very positive reactions, suggesting it is possible that Drug Risky could cure the terminal illness for some meaningful, but not large, subset of patients. In the first phase, there were also a non-negligible

number of side effects that were quite bad, especially as judged by the patient's previously established preferences. That is, the patient has been, in the past, quite adverse to certain side effects, and this drug did produce those side effects in the first phase of the trial. These are side effects that the doctor knows the patient has wanted to completely avoid in the past. In summary, there's a small chance that the drug could cure the patient, but also a small chance that it could cause side effects that the patient has been historically quite adverse to. The doctor feels that there is too much risk compared to the potential reward, and advises her patient to take Drug Safe. The patient though, upon hearing about both drugs, insists on Drug Risky.

This case provides a scenario where the prospects of death lead to a shift in the agent's preferences. Since Elster first introduced the concept of adaptive preferences, there has been a good bit of debate over the exact way to distinguish adaptive preference formation from character planning (Bovens 1992, 66-76, Colburn 2011, 67-70). In character planning, the agent purposely reacts to the change in circumstances by examining her preferences and coming up with a new preference ordering that fits her current circumstances (Elster 1983, 117-119, 138). While each method involves preferences adapting in response to changing circumstances, in adaptive preference formation preferences change through some non-conscious process that the agent neither controls nor is fully aware of. Character planning, on the other hand, is controlled by the agent. Yet, how we exactly pick out the agent being in control is quite difficult since the leading contenders (intentional vs. unintentional (Bovens 1992, 60-73, Colburn 2011, 56-57, Sandven 1999, 16-17), consciously vs. subconsciously made (Colburn 2011, 57, etc.) have met with problems in the literature. Fortunately, we don't need such precision in analysis to both understand that there is such a difference,[1] and that both types of preference change surely occur in life and death decisions.

That there is such a difference is made clear by the temporal difference that typically distinguish them in regular life.[2] Character planning involves a purposeful change of one's own preferences. But, of course, an agent cannot simply decide that her preferences ought to be altered and change them by fiat. Our preferences are not chosen in a moment and cannot be changed in a moment. Preferences are entrenched pieces of our psychology; they take work both to form

[1] Luc Bovens (1992, 59) argues that while there is a huge gray area where it is impossible to distinguish sour grapes and character planning, there are clearly instances where we can tell them apart.

[2] Bovens makes this point as well. Bovens argues that while there could be cases of sour grapes that take place over a long period of time, we wouldn't expect character planning to ever be immediate. In most cases, he believes that we can tell them apart by the length of time necessary for the change, and "a correct account of SG and CP should also provide an insight into what underlies this presumption" (Bovens 1992, 60, 77). Also see Tore Sandven's (1999, 24-29) discussion of mourning. Sandven argues that mourning necessarily takes time, but it seems to have the characteristics of an adaptive preference case.

and change. I cannot just decide to be fanatical about baseball after a lifetime of not really caring about the sport. I could though attempt to watch more games, attend a few, and learn more about the game – both in terms of how it works and what makes its fans love it – to slowly build in myself an appreciation of and preference for baseball fandom. Character planning typically, though not necessarily always, takes time because the purposeful changing of preferences can be difficult and time consuming.

Adaptive preference formation, on the other hand, will often be quick since the change occurs so automatically that the agent isn't likely to realize it happened. If a NFL season gets canceled and a diehard football fan finds herself suddenly preferring baseball (because, she has suddenly realized those football players are spoiled athletes), it is much more likely that her sudden preference change is an example of adaptive preference formation: the fan didn't do any work to make the change happen, but discovered it upon the moment of her need for it.

There is no guarantee that a quick change is adaptive preference formation and a long one is character planning. Further, to the agent who has a new or stronger preference, it may often be nearly impossible to determine, at least in the moment, that it wasn't purposely created or strengthened. Later, the agent may be able to tell that the preference wasn't the result of critical reflection or that it too easily subsided when circumstances changed back. Yet, when the change occurs, the preference will not appear to be foreign, but instead will seem to the agent to belong to her as much as any of her other preferences. But, as outsiders we can make an educated guess that a change resulted from adaptive preference formation if the change happened quicker than we would reasonably expect character planning to take. There is a difference between the two even if we lack a precise analysis of this difference.

It is likely that both adaptive preference formation and character planning occur in life and death situations. When a patient has a terminal disease, he is likely to have a large number of preferences that can no longer be satisfied. Any preference that requires the patient to live longer than her disease will allow are effectively ruled out. It is then in that patient's interests to figure out how to deal with all of those unfeasible preferences, either by giving them up, replacing them, or finding a way to satisfy them in a shorter time period. Like any circumstances that severely change an agent's feasible option set, achieving autonomy would require figuring out how to deal with the preferences affected by a terminal disease.

Some agents will react to this change in their option set with character planning, but surely others will have their preferences adapt unconsciously. While it makes sense to plan out your preferences once you realize you are going to die, not everyone will react calmly, deliberately, or rationally in response to death. For many people, the stark realization that death is near will trigger psychological reactions, including preferences unconsciously adapting so that the agent can still find some level of satisfaction prior to dying. Terminal diseases are likely to lead

to both character planning, which likely would be autonomous, and adaptive preference formation, which would show a lack of autonomy.

An agent who demands Drug Risky could fall under two possible characterizations. The agent could have come to his circumstances, which consists of a terminal disease and an aversion to certain side effects, with a conscious and purposeful decision to try to alter his preferences. When the doctor presents the patient with an option that could save his life, but that also could bring about the side effects that he abhors, the agent will work to make it so that he prefers the chance of being cured. If he succeeds, the patient could autonomously choose Drug Risky through character planning. The agent also could be unable to face his impending death calmly, and his psychological reaction may cloud the way he perceives his options. Such an agent could develop an adaptive preference to discount the positives of Drug Safe, as well as the negatives associated with Drug Risky. This patient would be choosing Drug Risky non-autonomously.

Adaptive preference formation, which can lead patients to make non-autonomous decisions, can cause tricky problems for medical professionals trying to respect their patients' autonomy. I will concentrate on the question of whether it is permissible for doctors to put patients in situations where preferences are likely to unintentionally adapt. In other words, our task is to figure out whether it is respectful of autonomy for a doctor to bring up a medical procedure or drug that they do not believe the patient can autonomously choose. This question is made especially tricky by the fact that it doesn't appear to be respectful of autonomy to hold back information. To better explore these issues, it will be necessary to pause and examine autonomy and respect.

7.3 Autonomy

It will be useful to develop a general sense of autonomy. There are, of course, a multitude of distinct autonomy theories,[3] but choosing and favoring a single one may stifle the ability to generalize the discussion's results. For this reason, I will attempt to give an autonomy account that incorporates some of the basic concepts that are found across many other autonomy accounts, while at the same time leaving open some theoretical issues that remain quite contentious in the autonomy literature.

"Autonomy" literally means self-legislating. To figure out what autonomy is, we must determine what it would mean for an agent to organize her actions according to legislation that she sets for herself. We can distinguish two general areas in the action process where an agent can develop and adhere to such legislation. First, an agent can determine what kind of values, desires, reasons, etc. she wants to be motivated to enact. The first stage of autonomy involves the

[3] See discussions on the various possible meanings of "autonomy," such as Arpaly (2004, 117-130). See also McKenna (2005, 206-207).

forming and shaping of the agent's motivational set. The second stage involves the agent figuring out how to act in particular circumstances so that she can satisfy her preferences as set in the first stage.

In both stages, the motivational set can provide the legislation needed to make sense of "autonomy." By the time agents reach an age of sufficient maturity to be considered autonomous, the agent's motivational set has largely been established in her childhood. When the agent looks to assess and possibly revise a piece of her motivational set, she will do so in light of how it fits with other pieces of her pre-existing set. Autonomy cannot require that an agent create a new motivational set from scratch as that would clearly be too much to ask for.[4] So, the original motivational set, largely formed in childhood, provides the initial legislation for the formation and shaping of new pieces, and the new pieces, endorsed through a greater amount of critical thinking, take their place in providing some more rationally grounded legislation for future pieces.

In the second stage of autonomy, the motivational set once again provides the legislation the agent needs to figure out how she ought to act. In most circumstances, it is simple for an agent to act according to her motivational set because there aren't any potential conflicts and the circumstances provide no difficulty in figuring out what ought to be done. In these circumstances, there is no need for substantial deliberation, or perhaps any deliberation. Immediately after waking up, the autonomous agent brushes her teeth without having to give it any thought because so much in her motivational set (the desire for white teeth, the value of healthy teeth, the need for avoiding smelly breath) is in favor of this action, and pretty much nothing is against it. Where circumstances are more complicated – where it is not clear how to act in the way that most satisfies what's important in her motivational set – the autonomous agent deliberates to determine what she should do. Her motivational set provides legislation for her deliberation, but this legislation derives from the agent, and she is the one holding herself accountable according to it. Thus, she is legislated, but only by her self.

This sketch of the two major stages of autonomy is rough. It is left open how the agent should determine what preferences should be in her motivational set, just as it is left open how the agent ought to deliberate in complicated circumstances. My points here are, first, that there is rational work involved in being autonomous, which can be separated into two main stages, and, second, we can have a basic understanding of autonomy that will be useful without having to settle the more contentious debates in the autonomy field.

---

[4] Joseph Raz (1986, 387-388) denies anyone believes in an "arbitrary self-creation doctrine," which would hold, "all value derives from choice which is itself not guided by value and is therefore free, i.e., arbitrary" (see also Feinberg (1986, 34), Friedman (2003, 8), Oshana (2005, 196-8), Noggle (2005)). For a scientific discussion of these sorts of issues, see Spence (2009). Charles Taylor though argues against Jean-Paul Sartre's view of radical choice of motives in a vein that makes it sound like it is such a view (Taylor 1976, 290-294).

7.4 Respecting and Disrespecting Autonomy

To analyze respect, it is useful to look at how we disrespect autonomy. I will examine two ways to disrespect autonomy. First, an agent can disrespect autonomy by interfering with the processes by which another person obtains and maintains her autonomy. Second, it is disrespectful of autonomy to put someone in a situation where the person will likely have to make a choice that either is unlikely to be autonomous or that would only be autonomous by default because the person is reacting as well as possible to the bad situation.

Let's start with disrespecting autonomy through interfering with autonomous processes. It would clearly disrespect autonomy to purposely or negligently interfere with the contents of an agent's motivational set or an agent's ability to critically think about how to act from that motivational set. Thus, brainwashing, propagandizing, slipping someone rationality inhibiting drugs, or taking away information that the person needs to deliberate over (either through deceit or withholding key facts the person has a right to know) are clear ways to undermine the first stage of autonomy.

This isn't to say that any influence on someone's motivational set or ability to think critically is disrespectful. We influence each other's motivational sets in various ways, almost always permissibly, all the time. When you try to convince a friend to try something new, when you openly display your preferences, or when you buy something for a friend that you think she will like, you are intentionally attempting to influence your friend's motivational set. There's an abundant supply of influences all around us: friends, family, colleagues, acquaintances, TV, movies, magazine, advertisements, Facebook, etc. Most of these sources, most of the time, permissibly influence our motivational sets.

Disrespecting autonomy involves influencing another person's autonomous processes in a way that attempts to take over or manipulate those processes. For an agent to be autonomous, she must have sufficient control over her processes of choosing and acting. When another person exerts too much influence, the agent's autonomy is undermined. Thus, even the attempt to place that much influence on another person disrespects that person's autonomy. This isn't to pinpoint the exact place where permissible influence turns into disrespect. It is sufficient that there are clear cases where influence becomes so strong that it counts as disrespectful.

Of course, we can also take over someone else's autonomous processes negligently.[5] It may be especially important in medical contexts to be aware of how we interact with people whose autonomous processes are vulnerable due to their diminished states. So, for example, when people are incredibly sick or confused, it is easier to negligently take over the person's processes by making decisions for them or pushing them towards decisions.

Insofar as disrespecting autonomy involves interfering with autonomous processes, respecting autonomy involves making sure influence doesn't become so

[5] Thanks to the editors of this volume for pointing this out.

great as to take over someone else's autonomous processes. For the most part, this merely involves giving people their space, not interfering with their rational processes, and being honest with them. Respecting autonomy in this sense simply requires allowing others to retain control of their autonomous processes.

In the second method, an agent disrespects autonomy by purposely or negligently putting another person in a situation where that person will either likely make a non-autonomous choice or will make a choice that only counts as autonomous by default because it is the best reaction to a bad situation. The former option involves placing someone in a situation where you know they aren't likely to remain autonomous, such as by tempting someone on a diet by bringing them to a bakery where you believe they cannot resist the sweets. Of course, this could also be done negligently if you bring a friend to a bakery who you should know was trying to stay on a diet. Though a person could autonomously choose to walk out of the bakery, it is disrespectful to put him in a situation where you believe he likely won't be able to make an autonomous choice.

The latter option must be included because we often place people in bad situations where they still are likely to make autonomous choices by reacting well to those bad situations. It would seem to be disrespectful to preclude the best option from someone's feasible set, but that person can still choose the second best option. Autonomy involves agents reacting to the situations they are presented with. If a person chooses what is now the best option available in her feasible set, she should still count as autonomous. It may be quite difficult to completely undermine autonomy without interfering with autonomous processes. It is up to the agent how she wants to respond to her circumstances; she may always find an autonomous response.

While brainwashing undermines autonomy, it is not clear that coercion does.[6] Here, I am using "coercion" to mean the use of a threat to make another person meet a demand. Of course, there is a great deal of debate over how we define "coercion" and whether it ought to include threats,[7] but here we can simply stipulate that we are specifically discussing coercion as it involves a threat. In a typical coercion case, such as a mugging, it is up to the victim to process the situation and determine what his best action is to meet his autonomously formed preferences. When the mugger says, "Your money or your life," it certainly can be an autonomous choice for the victim to give over his money since that action best

[6] For those who believe coercion and autonomy are inconsistent, see Beauchamp (2005, 311, 313, 315), Oshana (2003, 100-103), Ekstrom (2005, 151, 154), Mele (2001, 187). For a view that they may be consistent, see Ekstrom (2005, 145). And for views that leave it an open question, see Mele (2001, 185-6), Ekstrom (2005, 145), Noggle (2005, 87-88), Arpaly (2005, 163).

[7] For those who believe threats are necessary for coercion, see Nozick (1969, 447-453, 458-465), Raz (1986, 150), Bayles (1972, 21-23), Gert (1972, 30, 34-37, 47). For those who believe in coercive offers, see Zimmerman (1981, 131), Frankfurt (1988, 41-42), Feinberg (1986, 216-268), O'Neill (2000, 85-87, 94-95), Carr (1988, 65), Stevens (1988), Held (1972, 51, 54-57).

fits his autonomously arranged motivational set. The mugger has not prevented an autonomous choice since the victim chooses his best action given his circumstances of being mugged.

We can still say that the mugger has disrespected the victim's autonomy since she forced her victim into a situation where she knew he would have to make a choice that wouldn't be autonomously chosen without her interference. Let's say in such a situation that the mugger forces the victim into a "less preferable action." We can judge whether an action is more or less preferable through the agent's autonomously formed motivational set. Action A is preferable to action B if the agent's preferences would recommend A over B in any circumstances where both were available and no other preference was strongly influencing the choice. The agent would only choose B if it were consistent with A, A were not available, or some third preference would also be satisfied by choosing B.

We cannot determine whether the changing of another person's action is respectful simply by learning whether the new action is autonomous. After all, you could eliminate a person's preferable action (A), leaving him to perform his less preferable action (B). Instead, we ought to judge respect through whether the person is being changed to a preferable action. To change someone's action respectfully, it is necessary to move the person to an action that is at least equally preferable. The mugger disrespects autonomy by moving her victim into a situation where it is only reasonable to believe he would be forced to take a less preferable action, as judged by what it is reasonable to assume his motivational set is like. Muggers attempt to get their victims to choose actions that could be autonomously chosen, but aren't likely to be preferable since almost everyone would prefer walking away over losing their money or their lives. Thus, we can see the disrespect of autonomy in the elimination of preferable options for the purpose of making the other person choose a less preferable action.

That leaves the possibility of changing someone's autonomous action respectfully provided that the change is to a preferable action. This may seem like a high burden, but we do it all the time in ordinary interactions. If you want to change someone's choice, you can offer an enticement, you can try to convince the person that the change is actually preferable for her too, or you can point out that your relationship – friendship, family, colleague, or whatever – makes helping you a preferable action. When we respectfully try to change someone's mind, we tend to offer reasons, payments, gifts, or refer to the strength of our relationships. These enticements either point to why the change is already preferable or try to make the change preferable for the other person, and this shows our respect for the other person's autonomy.

We have two methods for disrespecting autonomy, which each can occur intentionally or negligently, along with two corresponding methods for respect. The first method of disrespect involved interfering with the processes that make up autonomous choosing. The corresponding method of respect involves not interfering with those processes or actually aiding them. The second method of disrespect involves either putting someone in a situation where he isn't likely to be autonomous or blocking someone's preferable action so that they must choose an

action that will be less preferable. Respecting autonomy either involves making sure that you don't tempt people into doing things that won't be autonomous for them or changing people's autonomous actions only if that can be done in a way that is preferable for them.

7.5 Respect and Adaptive Preferences in Terminal Disease Clinical Trials

We can now return to the medical arena to better understand the roles that respect and disrespect play in engaging with a patient with a possible adaptive preference. Given the two methods of respect and disrespect we have looked at, we should be able to figure out whether medical professionals have specific duties of respecting autonomy in such situations.

Let's remind ourselves of the scenario at hand. The patient wouldn't normally risk Drug Risky because of its potentially damaging side effects, but he does choose Drug Risky over Drug Safe in this situation since Drug Safe provides no chance of curing his terminal disease. The doctor knows the patient wouldn't normally risk these side effects. Of course, this situation is unique since the patient hadn't previously had a terminal disease.

If the doctor has good reason to believe the patient will react with character planning, then the doctor certainly has a moral obligation to inform the patient about Drug Risky. If the doctor really can't be sure whether the patient will react with an adaptive preference formation or with character planning, the doctor's responsibility is almost surely to give the patient the option of Drug Risky along with encouragement for the patient to take his time to decide rationally in light of the possible risks.

Things become trickier when the doctor has good reason to believe that the patient won't engage in character planning. She may realize that the patient has neither the time nor the emotional calmness to react with character planning. His emotional state since learning about his condition makes it clear that he isn't likely to make his decision through calm deliberation. The doctor believes that a preference for Drug Risky would likely be an adaptive preference, and would not reflect an autonomously made choice. This doctor is faced with a moral dilemma since she believes she should inform her patients of their options, but she also values helping her patients avoid choices that won't be made autonomously.

The problem is that each of the doctor's available options seems to disrespect autonomy. First, a failure to provide information about available options can interrupt the autonomous process of making a choice. An agent who deliberates with inadequate information may choose in a way that is procedurally rational – there is nothing inconsistent in his reasoning – but it would be less substantially rational in that the agent is missing key facts that he is trying to deliberate over. Withholding key information is a method of disrespecting autonomy due to its interference with autonomous processes.

In the case at hand, it would appear to be disrespectful of the patient's autonomy to withhold the option of Drug Risky, which has a small chance of curing the patient's disease. Even if there is reason to think the patient would choose Drug Risky non-autonomously, it certainly could be autonomously chosen, theoretically speaking. It would seem to be disrespectful not to give the patient the chance to determine for himself if he wants to take the risk.

On the other hand, it can be disrespectful of autonomy to put someone in a situation where you believe he will not make an autonomous choice. In this case, the doctor believes that the patient's autonomous choice would be to take Drug Safe and work towards acceptance of his terminal condition. Adding Drug Risky into his choice set would probably lead him to develop an adaptive preference for that drug. The doctor believes that informing the patient about the clinical trial will likely lead the patient to take a less preferable, and possibly even non-autonomous, choice, as judged by the patient's motivational set.

Worse, the patient will lose the highly valuable effect of Drug Safe that it will allow him to work on accepting his death calmly with less suffering. By introducing Drug Risky into the equation, even if the patient chooses Drug Safe, it will be harder to accept his death since he knows he turned down a chance at a cure. By adding Drug Risky as an option, the doctor risks the patient's peace of mind.

The doctor appears to be locked into a catch-22: both choices seem to disrespect her patient's autonomy. We ought to wonder whether there is a basis to prioritize one kind of respect in this scenario. It seems to be less disrespectful to put someone in a bad situation than it is to withhold information especially if the doctor doesn't intend to put the patient into a bad situation. The doctor must put the patient into a bad situation because the doctor can't withhold pertinent information. It is because the doctor wants to respect the patient's autonomy by providing relevant information that the bad situation becomes an unintended byproduct of the doctor's action. Thus, we could say that it is acceptable to place a patient in a situation where his autonomy may be compromised as long as the doctor only does so to respect the patient's autonomy in some other fashion.

In this case, we also have a justification for withholding information. The doctor can withhold information about a certain risky treatment plan, such as with Drug Risky, where she thinks telling the patient would disrespect his autonomy by putting the patient in a situation where he may get his hopes falsely up, become exposed to too much risk, or choose against his best interests, as judged by his motivational set. In such a case, it does seem the doctor can justifiably withhold information because the information is not necessary for the patient to make an autonomous choice.

There are two viable, and yet conflicting, conclusions in this case where the doctor doesn't think character planning is likely. We can and should live with this tension. The first argument indicates that the doctor could be justified in providing information about Drug Risky because she feels the ultimate decision should be left in the patient's hands, even if the decision may turn out to be non-autonomous. The second argument indicates that the doctor could be justified in

withholding the information because she feels that Drug Risky is wrong for her patient in her expert opinion.

We should live with this tension because real life moral dilemmas seldom simplify to a single answer. We cannot know in advance how a particular medical consultation should go since every case has a large number of variable parameters. The arguments presented here provide allowances for the doctor. The particular situation may make one action obligatory and the other impermissible, but there is no reason to think one answer is obligatory across all situations.

It then ought to be left up to the medical professional, given these guidelines. Where there is a chance that the patient will react with character planning, and where the doctor's expertise doesn't rule out the solution as too ineffective or too risky, then the doctor should tell the patient about Drug Risky. Where the chances of character planning are non-existent or where the doctor's expertise suggests that the patient cannot autonomously choose the drug in question, then the doctor should withhold the Drug Risky information. Provided that doctors know their patients sufficiently well and are striving to figure out the best way to respect patient autonomy, there are multiple permissible methods to respect autonomy in these cases.

7.6 Conclusion

We have seen why life and death medical situations could bring about adaptive preferences, which may undermine the autonomy of patients. This result can be very problematic since it is well accepted that medical professionals ought to respect their patients' autonomy. It is then clearly important for medical professionals to be aware of how adaptive preferences may come up in medical contexts, and to consider how they can respect autonomy in such situations.

To focus our attention on this issue, we examined a clinical trial where a life and death situation could lead to an adaptive preference that undermined the patient's autonomy. A patient has a terminal illness, for which there is a regularly prescribed drug, Drug Safe, that will ease the patient's suffering, briefly extend his life, and hopefully allow the patient to come to grips with death. There is also a clinical trial for another drug, Drug Risky, which has had, in its first phase, some non-negligible number of worrisome side effects, but also a small number of very positive results that indicate it could cure the disease for a small number of patients.

We imagined a patient that would choose Drug Risky. There are a few ways to interpret that choice. The patient could have changed her motivational set, through character planning, so that Drug Risky is her autonomous choice. Character planning typically involves gradual changes since we cannot alter our preferences by fiat but must instead work to bring about motivational changes. So this patient's alteration could take awhile, requiring time for the patient to make his decision. Of course, this could be a rare case where character planning happens quickly – right in the doctor's office – in which case it is possible that a quick choice could be autonomous. The patient also could have engaged in character planning between learning about the terminal disease and going to the doctor to

hear about treatment options. Thus, there are scenarios where the patient autonomously chooses Drug Risky right in the doctor's office, but they are less likely, and we would expect a preference for Drug Risky that occurs immediately to be a result of the quicker adaptive preference formation. The final way to interpret the patient's choice then is that it is based in the patient's adaptive preference. In that case, the choice would not be autonomous.

Given these possibilities, the doctor will have a range of reactions and predictions, depending on how well she knows the patient. Whether the doctor reacts respectfully depends on how the patient is likely to respond, how well the doctor knows that, and on the kind of respect that ought to be prioritized in the situation.

We saw that there were two distinct, though possibly overlapping, ways to respect autonomy that correspond to two distinct ways to disrespect it. The first method of disrespect involves interfering with someone's autonomous processes, such as by attempting to brainwash the person or trying to inhibit the person's rationality, either through rationally inhibiting drugs or inhibiting the person's being adequately informed. The corresponding method of respect involves either leaving someone's autonomous processes intact or aiding the person by making him a better critical thinker, aware of more options, or more adequately informed about the relevant facts for his deliberation.

The second method of disrespect involves putting someone in a position where either he will be unlikely to be autonomous or will only be autonomous by default because he must choose a less preferable option. The corresponding method of respect involves either not changing people's possibly autonomous choices or only changing possibly autonomous choices if you can move the person to an equally or more preferable action, as judged by his motivational set.

The adaptive preference cases present tricky issues for the doctor because these methods of respect appear to conflict. On the one hand, a doctor fails the first method if she withholds pertinent information from her patient. On the other hand, the doctor fails the second method if she places her patient in a situation where he is unlikely to make a preferable autonomous choice. The ethical doctor who cares about respecting patient autonomy finds herself stuck in a catch-22.

Fortunately, I was able to argue that the conflict can be resolved in different ways in different situations: some situations require the doctor to provide the information and others require withholding it. There is no single answer that applies to all relevant cases, but the conflict is resolvable in each case. If the doctor believes that the patient has a chance of engaging in character planning and recommending Drug Risky does not violate her professional judgment, then the doctor can respectfully give that option. If the doctor believes the patient almost certainly will not react autonomously or it is clear that her professional expertise makes it impossible to recommend Drug Risky for this patient, then the doctor can respectfully withhold the information. While adaptive preferences present problems for ethical doctors who seek to respect autonomy, there are methods available to deal with situations that are likely to bring about adaptive preferences while still respecting patient autonomy.

References

Arpaly, N.: Responsibility, Applied Ethics, and Complex Autonomy Theories. In: Taylor, J.S. (ed.) Personal Autonomy: New Essays in Personal Autonomy and its Role in Contemporary Moral Philosophy, pp. 162–180. Cambridge University Press, Cambridge (2005)

Arpaly, N.: Unprincipled Virtue. Oxford University Press, New York (2004)

Bayles, M.: A Concept of Coercion. In: Pennock, J., Chapman, J. (eds.) Coercion: Nomos XIV, pp. 16–29. Aldine Press, Chicago (1972)

Beauchamp, T.L.: Who Deserves Autonomy and Whose Autonomy Deserves Respect? In: Taylor, J.S. (ed.) Personal Autonomy: New Essays in Personal Autonomy and its Role in Contemporary Moral Philosophy, pp. 310–329. Cambridge University Press, Cambridge (2005)

Bovens, L.: Sour Grapes and Character Planning. J. Philos 89, 57–78 (1992)

Carr, C.C.: Coercion and Freedom. Am. Philos Quart. 25, 59–67 (1988)

Colburn, B.: Autonomy and Adaptive Preferences. Utilitas 23, 52–71 (2011)

Ekstrom, L.W.: Autonomy and Personal Integration. In: Taylor, J.S. (ed.) Personal Autonomy: New Essays in Personal Autonomy and its Role in Contemporary Moral Philosophy, pp. 143–161. Cambridge University Press, Cambridge (2005)

Elster, J.: Sour Grapes: Studies in the Subversion of Rationality. Cambridge University Press, Cambridge (1983)

Feinberg, J.: The Moral Limits of the Criminal Law, vol. 3 Harm to Self. Oxford University Press, New York (1986)

Frankfurt, H.: Coercion and Moral Responsibility. In: Frankfurt, H. (ed.) The Importance of What We Care About, pp. 26–46. Cambridge University Press, New York (1988)

Friedman, M.: Autonomy, Gender, Politics. Oxford University Press, New York (2003)

Gert, B.: Coercion and Freedom. In: Pennock, J., Chapman, J. (eds.) Coercion: Nomos XIV, pp. 30–48. Aldine Press, Chicago (1972)

Held, V.: Coercion and Coercive Offers. In: Pennock, J., Chapman, J. (eds.) Coercion: Nomos XIV, pp. 49–62. Aldine Press, Chicago (1972)

McKenna, M.: The Relationship between Autonomous and Morally Responsible Agency. In: Taylor, J.S. (ed.) Personal Autonomy: New Essays in Personal Autonomy and its Role in Contemporary Moral Philosophy, pp. 205–234. Cambridge University Press, Cambridge (2005)

Mele, A.: Autonomous Agents. Oxford University Press, New York (2001)

Noggle, R.: Autonomy and the Paradox of Self-Creation. In: Taylor, J.S. (ed.) Personal Autonomy: New Essays in Personal Autonomy and its Role in Contemporary Moral Philosophy, pp. 87–108. Cambridge University Press, Cambridge (2005)

Nozick, R.: Coercion. In: Morgenbesser, S., Suppes, P., White, M. (eds.) Philosophy, Science, and Method: Essays in Honor of Ernest Nagel, pp. 440–472. St. Martin's Press, New York (1969)

O'Neill, O.: Bounds of Justice. Cambridge University Press, Cambridge (2000)

Oshana, M.: How Much Should We Value Autonomy? In: Paul, E.F., Miller, F.D., Paul, J. (eds.) Autonomy, pp. 99–126. Press Syndicate of the University of Cambridge, Cambridge (2003)

Oshana, M.: Autonomy and Free Agency. In: Taylor, J.S. (ed.) Personal Autonomy: New Essays in Personal Autonomy and its Role in Contemporary Moral Philosophy, pp. 183–204. Cambridge University Press, Cambridge (2005)

Raz, J.: The Morality of Freedom. Oxford University Press, Oxford (1986)
Sandven, T.: Autonomy, Adaptation, and Rationality: A Critical Discussion of Jon Elster's Concept of "Sour Grapes," Part I. Philos Soc. Sci. 29, 3–31 (1999a)
Spence, S.: The Actor's Brain. Oxford University Press, New York (2009)
Stevens, R.: Coercive Offers. Aust J. Philos 66, 83–94 (1988)
Taylor, C.: Responsibility for Self. In: Rorty, A.O. (ed.) The Identities of Persons, pp. 281–300. University of California Press, Berkeley (1976)
Zimmerman, D.: Coercive Wage Offers. Philos Pub. Aff. 10, 121–145 (1981)

Chapter 8
Preference Adaptation and Human Enhancement: Reflections on Autonomy and Well-Being

Maartje Schermer

> The world's stable now. People are happy; they get what they want, and they never want what they can't get.
> (Huxley, *Brave New World*)

8.1 Introduction

In Huxley's well-known novel *Brave New World* people are engineered in such a way that they cannot help but wanting the things they have, and aspiring the social positions they are in fact designed to hold. In a sense, one could call this a form of human enhancement – the people in this Brave New World are all very content and happy. However, as the novel and the above quotation make tangible: wanting only those things you can have may make the world more stable and individuals happier, but it also makes people lose an important part of their humanness. What they may gain in happiness, they lose in self-determination and autonomy. Automatically adapting their wants to their possibilities without any reflection, the people in Brave New World are exemplars of the problem of adaptive preferences.

In this chapter I will explore the problem of preference adaptation in relation to the debate on human enhancement, i.e. the debate on emerging biomedical

Maartje Schermer
Department of Medical Ethics and Philosophy
Erasmus MC University Medical Center
Rotterdam
The Netherlands
e-mail: m.schermer@erasmusmc.nl

J. Räikkä and J. Varelius (Eds.): *Adaptation and Autonomy*, SAPERE 10, pp. 117–136.
DOI: 10.1007/978-3-642-38376-2_8 © Springer-Verlag Berlin Heidelberg 2013

technologies that aim to improve human capacities and traits, like cognition, physical strength, beauty or lifespan. Adaptive preferences are preferences that we have formed unconsciously, as an adaptation to the situation that we find ourselves in, in order to avoid the pain and suffering associated with the dissonance between our actual situation and the situation we would prefer. Adaptive preferences and similar forms of adaptation of wants, hopes and desires in light of actual or perceived options have been criticised in the ethics literature for being non-autonomous or not rational. Moreover, the judgements that people make regarding their own happiness or well-being have been criticised as failing to reflect their true well-being, if they are based on adaptive preferences or other forms of adaptation to unfavourable situations. If preferences regarding enhancement turn out to be adaptive preferences, this could mean, therefore, that such preferences are non-autonomous and/or do not reflect our true well-being. In what follows, I will examine this possibility and its consequences in further depth.

First, the notions of adaptation and adaptive preferences will be discussed, and next, it will be argued that some preferences with regard to human enhancement – both preferences in favour and against enhancement – can be considered 'adaptive preferences.' The main example of human enhancement technology will be life extension, but the argument would hold as well for other examples of enhancement, such as cognitive enhancement or physical enhancements. Secondly, I will argue that this implies that not all preferences with regard to enhancement – and especially life extension – are autonomous preferences, but that it is possible to transform non-autonomous preferences into autonomous ones. The enhancement debate can profit from this insight since it is important that preferences, whether they are in favour of or against enhancement, are autonomous. Finally, the impact of adaptive preferences on well-being and happiness will be briefly evaluated. The notion of adaptation is crucial in many discussions on well-being, especially when linked to the capability approach on well-being. I will argue that the enhancement debate could profit from some of the insights gained in this field.

8.2 Sour Grapes and Other Worries about Preference Adaptation

8.2.1 Sour Grapes and Character Planning

Since Jon Elster's discussion of the problem in his 1983 Sour Grapes, the problem of adaptive preferences has also been known as the sour grapes problem. As readers of this volume will be aware of, Elster refers to La Fontaine's (originally Aesop's) fable of the Fox and the Grapes to explain what he means by preference adaptation. The hungry fox saw some delicious grapes hanging high on a vine, but however high he tried to jump, they remained beyond his reach. Noticing this, the fox gave up his attempt to get the grapes, concluding that he did not want them anyway because they were probably sour.

The point is that the fox unconsciously adapts his preference for the grapes when he finds he cannot have them; according to Elster this unconscious U-turn in the fox's preferences for the grapes is therefore irrational. Preferences are rational, on Elster's account, if they are acquired in a rational way, and they are so only if they are autonomous, i.e. intentional. Therefore, unintentional and unconscious adaptations of preferences such as those of the fox, are irrational and not autonomous. Adaptive preference changes take place 'behind the back' of the person concerned (Elster 1983, 117); they are not the result of intentional preference formation in response to reasons or desires, but simply 'caused' by the change in the person's set of options. For these reasons, Elster finds sour grape adaptation highly problematic. Although by sour grapes adaptation one gains in well-being because ones preferences can be satisfied easier, one loses autonomy. For Elster this trade-off is undesirable.

In Elster's work, the notion of adaptive preferences refers to the unintentional adaptation of our preferences to our options; to the adjustment of our wants to our possibilities. Other authors, however, have used the notion of adaptive preferences more broadly, to include all kinds of adaptations of wants, aspirations, hopes, desires and preferences to the situation one finds oneself in; and to include adaptations that can be conscious or intentional as well. The various notions of adaptation and related phenomena that influence mental reaction, preference formation and human behaviour have been nicely mapped out by Clark (2012). These include, e.g. social conditioning and indoctrination, hedonic adaptation, preference or value formation through learning and experience and false expectations and misguided optimism.

Elster himself has also discussed a number of related forms of preference adaptation. In doing so, he has made one especially important distinction between the unconscious causal adaptation of preferences to one's options – adaptive preferences – and the conscious, deliberate adaptation of one's preferences to one's feasible option set, which he calls Character Planning. Character Planning implies that a person intentionally adapts and shapes his own desires and preferences to fit his limitations and feasible options; it is therefore a much more autonomous way of adapting to ones circumstances than Sour Grapes style adaptation. Elster even concedes that: "there is a respectable and I believe valid doctrine that explains freedom in terms of the ability to accept and embrace the inevitable" (Elster 1983, 119). Another difference is that while adaptive preferences, according to Elster, typically take the form of downgrading inaccessible options, deliberate character planning tends to upgrade the accessible ones.

Although the difference between Sour Grapes (SG) adaptation and Character Planning (CP) adaptation is crucial in terms of autonomy of the resulting preferences, the distinction may be less clear than Elster presupposes. According to Luc Bovens, there is a broad grey area between both phenomena: "for a great many cases of preference adjustment in reference to the feasible alternatives, it is not clear whether to categorize them as instances of SG or as instances of CP" (Bovens 1992, 59). Bovens discusses an alternative to Elster's theory to

distinguish between the two. Other authors as well have tried to come up with a more satisfying account of the difference between SG and CP, and the consequences for the autonomy of these preferences (e.g. Colburn 2011). This issue will be taken up further in section 8.4.

8.2.2 Adaptation and Well-Being

The notion of adaptation also plays a crucial role in a somewhat different field of scholarly work, namely that of capability theory and human development and well-being. Authors in this field are not mainly concerned with the rationality or autonomy of adaptive preferences, but with the relationship between adaptation and subjective evaluations of well-being. Capability-theorists like Amartya Sen and Martha Nussbaum point out that people under circumstances of poverty, deprivation and injustice can still consider themselves to be happy or well-off. They contend that desires and attitudes are malleable and can be adapted to the situation in which people live, but that because of this phenomenon of adaptation their subjective assessment of their situation does not really reflect the truth about their well-being (Teschl and Comim 2005, Clark 2012).

The notion of adaptive preferences explains why people in very poor and deprived circumstances or in situations of social oppression report subjective well-being, while objectively – as seen from the outside – their well-being is low. The Happy Slave is the paradigmatic example of this predicament. Although he is poor and has no freedom, he says he is happy simply because he cannot imagine any other way of life for himself. A similar mechanism makes for example poor women in India acquiesce in their position of subordination and abuse (Nussbaum 2000).

8.2.3 Healthcare and the Disability Paradox

Finally, it is worthwhile to point out that adaptive preferences are also a well-known phenomenon in medicine and healthcare (Barnes 2009). Here, the problem is known as the ‘disability paradox’ and refers to the adaptation of a patient’s aspirations and expectations to her situation of disability or chronic illness, in such a way that her subjective well-being remains intact. This paradox was first described by Albrecht and Devlieger, who formulated it as follows: “Why do many people with serious and persistent disabilities report that they experience a good or excellent quality of life when to most external observers these individuals seem to live an undesirable daily existence?” (Albrecht and Devlieger, 1999) The observation that disabled or ill persons who are in the eyes of many healthy people ‘badly off’ still report subjective well-being poses a problem for quality of life measurements in healthcare and health research. What should be taken to represent the truth about these patients’ well-being or quality of life: their subjective reports or some objective measures?

Worries about sour grapes and happy slaves suggest that objective measures should be used, because subjective reports are probably based on non-autonomous or irrational preferences adaptation. One objection to this view could be that it disregards the possibility that adapting one's preferences to ones situation can be a wise thing to do, a sign of resilience. The term resilience refers to "a pattern of functioning indicative of *positive adaptation* in the context of significant risk or adversity" (Ong et al. 2009) and to "the capacity to maintain, or regain psychological well-being in the face of a challenge" (Ryff et al. 2012). Adaptation to inevitable hardship and circumstances can be a vital and conscious strategy. Elster's distinction between adaptive preferences and character planning is important here and can help to conceptualise different forms of the disability paradox: those involving resignation and those involving resilience.

8.2.4 The Fisherman's Wife: Upward Adaptation

The discussion on adaptation has focused on the sour grapes problem: adaptation in light of a lack of options, or the closing down of options. However, a similar adaptation of our wants and desires to our options and possibilities can also occur with the opening of new options or in light of an abundance of options. It is known from many psychology and economic studies that "people adjust aspirations upwards in light of previous achievements or social comparison with others" (Clark 2012, 72). This is known as *upward adaptation*.

As a parallel to the fable of the Fox, consider the story of the brothers Grimm about the Fisherman and His Wife. The fisherman and his wife lived in a shack and had just enough to live a decent life. When one day the fisherman caught a fish that turned out to be an enchanted prince, and that could grant him every wish, the fisherman asked for a cottage for his wife. However, soon the wife was not satisfied with the new cottage anymore and asked for a villa, and when she became used to that luxury she asked a palace, and so on..... Of course this story does not end well. But it shows that preferences can be adapted not only downward, but upward as well. Once the option of endless wish-fulfilment is opened up to the fisherman's wife, she becomes dissatisfied with what she had before, and wants ever more.

This mechanism is known in psychology and economics as the hedonic treadmill or hedonic adaptation. This is also a form of adaptive preferences, although there are dissimilarities with Elster's Sour Grapes. First, upward adaptation need not necessarily be unconscious and it need not be irrational; but to the extent that it is, it shares an important characteristic with Sour Grapes adaptation. Second, in the case of Sour Grapes the downward adaptation of one's desires has a function of taking away the pain and frustration of not being able to have what one wants. Adaptation in that case yields the benefit of relief and increased subjective well-being. This is not the case with upward adaptation. On the contrary, upwards adaptation renews the frustration of not being able to get what one wants, by expanding one's preferences until they cannot be satisfied

anymore.[1] However, this appears to be equally irrational as unconscious downward adaptation. The important point here is that the impact of unconscious and irrational forms of adaptation on the authenticity of a person's desires and on his autonomy is similar, regardless of the direction of the adaptation, as will be discussed more extensively in section 8.4.

8.3 Human Enhancement and Adaptive Preferences

8.3.1 Preferences Regarding Human Enhancement

How does the notion of preference adaptation play a role in the debate on human enhancement? Human enhancement can be defined in various ways, but one definition is that "a biomedical enhancement is a deliberate intervention, applying biomedical science, which aims to improve an existing capacity that most or all human beings typically have, or to create a new capacity, by acting directly on the body or the brain" (Buchanan 2011, 23). Examples of such enhancements include cosmetic surgery, doping in sports, cognitive enhancement by way of psycho pharmaceuticals, and life-extension by way of various biomedical interventions. Some of these enhancement technologies and options are with us already; others are more speculative and futuristic, or still in early phases of development. Regardless of the actual current possibilities for enhancing human traits and functions, however, it is important to note the way of thinking that underlies the enhancement movement. This way of thinking can be characterised as optimistic, with a strong belief in the powers and possibilities of human beings and with a great faith that innovation will lead to progress.

Proponents and enthusiast about human enhancement promise that biotechnological enhancements will free people from the limitations of human nature, will greatly improve their capacities and their options, and will thereby ultimately improve their lives. Especially the transhumanist movement voices high expectations. As Nick Bostrom says, "[they] hope that by responsible use of science, technology and other rational means we shall eventually manage to become post-human, beings with vastly greater capacities than present human beings have" (Bostrom 2003, 493).

Enhancement technologies open up – or at least they promise to open up – a whole new range of opportunities and options. Suppose we could enhance our cognition, for example; this would enable us to learn faster, get more work done in smaller amounts of time or with less effort, and it might even enable us to understand things or do things we cannot do right now. Or consider the possibility of life extension. A lot of work is currently being done in the field of biogerontology and attempts to slow down, stop, or even reverse the aging process are underway. Even if the primary goal is not to extend human lifespan but merely to prolong the *healthy* lifespan, as most scientists in the field claim they aim at, the

[1] I owe this point to Joel Anderson.

side-effect of healthy aging may be a the prolongation of lives. Extension of the human lifespan to about 120 years is deemed a realistic prospect by many experts (Schermer and Pinxten 2013). Some even claim that longer lifespans may be feasible, and the most optimistic proponents of prolongevity such as Aubrey de Grey are confident we may eventually reach a state of near immortality. The prospect of longer lives may seem very attractive to many and the mere promise of such an enhancement may spark a desire to actually live up to 120, or 150, or even more and to support the quest for lifespan extension.

As the example of lifespan extension already shows, desires or preferences for enhancement can take various shapes. First, we can distinguish between individual desires and 'collective desires'. An individual can have a desire or a preference to use a certain specific enhancement technology. For example, a woman may have a desire to improve her looks by way of cosmetic surgery. A professional cyclist may want to improve his endurance by using a specific performance enhancing drug. Collective desires on the other hand are more like the 'hopes and dreams' of larger groups of people, even of a whole community or society. They explain why we undertake collective action and work on projects to change and improve our existing situation. Our collective desire to live longer, for example, inspires much research and development in genetics, biogerontology, biotechnology, nanotechnology et cetera. Many people want to find the 'fountain of youth' and so the preference to live longer lives makes us pursue specific lines of research and technology development (Post and Binstock 2004). This desire is not directed at having or using a concrete single available technology but it is a desire that sets an aim for collective action. Since in the end a collective preference is the sum of all individual preferences,[2] and since applying the notions of adaptation to collective preferences poses many theoretical complexities, the remainder of this chapter will focus on individual preferences and consider collective preferences to be a derivative of individual preferences.

Second, while some preferences with regard to enhancement favour the use of enhancement technologies, others oppose them. One can have a preference *not* to use enhancements. For example, a sportsman can prefer to rely on his own strength, endurance and training, and reject the use of doping as unnatural or unfair. Or a woman can reject the use of cosmetic surgery because she feels it sets sexist norms, or because she prefers a 'natural' look. Likewise, we can reject the further development of enhancement techniques and oppose, for example the quest for significant extension of the human life span, or even for eternal life. In the ethical discussion on lifespan extension, both positions for and against have been argued for. As Underwood et al. note, "advocates of strong life extension technologies have assumed a high or even near universal demand for such

[2] Perhaps a collective preference is more than the sum of individual preferences, since those preferences reinforce and influence each other. Especially in political philosophy and philosophy of economics the question what exactly a 'collective preference' is, is disputed (e.g. Gilbert 2001). It would take too far to explore this issue here in further depth, however.

technologies. Opponents suggest that the community would find life extension inherently undesirable" (Underwood et al. 2009, 496). Research among the general public shows that there exists a diversity of opinions and preferences regarding life extension. While some would welcome the opportunity to extend their lives, others would not even entertain the possibility (Underwood et al. 2009). In a study by Partridge et al. (2009) some people expressed a (strong) desire for life extension techniques, while others opposed or were ambivalent. Participants who supported the research efforts to slow ageing and increase the maximum lifespan were mostly also interested in using such (as of yet hypothetical) technologies themselves. Others opposed on religious grounds – "the lives we have is how God meant us to be" – or on similar grounds, expressed as a concern for naturalness: "it seems totally unnatural. It seems to be upsetting the natural sequence of things" (Partridge et al. 2009, 72).

8.3.2 Are Enhancement Preferences Adaptive?

We have seen that individual preferences with regard to enhancement can be preferences *for* or *against* using enhancement technologies; or for or against the development of enhancement technologies. In light of the discussion on adaptive preferences, we can ask whether some of these preferences with regard to enhancement are perhaps adaptive preferences. On closer inspection, there seem to be two ways in which enhancement-related desires or preferences can be adaptive.

First, the rejection of enhancement technologies may be based on adaptive preferences. We are so used to *not* being able to live for 150 years – let alone forever, – or to enhance our cognition by a simple pill, that we do not even desire it. As long as we take our current situation as a frame of reference and as long as we do not consider enhancement a realistic option, we may not form a desire for it. We may claim that we are good enough as it is; that this is our normal and natural state and that we are happy to be as we are. But wouldn't that be a classical case of Sour Grapes? Wouldn't it be like the fox who says he does not want to live forever anyway, because he knows that he can't? We could argue that such satisfaction with the state-of-affairs is non-autonomous, because we have unconsciously adapted to the limits our nature poses.[3] Indeed, it has been suggested by pro-enhancement authors that opponents of enhancement do suffer from a form of adaptive preferences. For example, Bostrom and Ord (2006) argue that rejection of enhancement is often based on status quo bias and that it is therefore irrational.

Second, it might also be the other way around: preferences in favour of enhancement can be adaptive preferences. We may only want cosmetic surgery,

[3] A problem with this line of argument is that it threatens to turn everything we do not desire, everything we are satisfied with, into a form of adaptive preference. Our satisfaction with any current state of affairs could be called 'adaptive' if we had reasons to prefer another state of affairs but don't prefer it because we believe it to be unreachable.

performance enhancing drugs, or life-extension because it is available, or because it promises to be available. This is a classic example of 'supply creates demand'. We may only want these things because they are available to us, or because they promise to be available if we only invest enough effort, and money, in research. This is not a classical example of Sour Grapes adaptive preferences, but of upward adaptation or Fisherman's Wife adaptation. Nevertheless, this way of preference formation may impact on autonomy the same way as Sour Grapes adaptation does, because, as Elster argues, what matters is that the valuation is determined by the available options, or by the options perceived to be available. This point will be further developed in section 8.4.2. Finally, to avoid any misunderstanding, I would like to state clearly that my claim here is not that preferences with regard to enhancement are always necessarily adaptive preferences. The point is merely to suggest that in some cases, they can be. The next question to be considered is what this means for the autonomy of such preferences.

8.4 Enhancement, Adaptation, and Autonomy

If at least some preferences regarding enhancement are adaptive in nature, this poses two distinct problems. In line with capability theorists, one could ask whether preferences regarding enhancement do really reflect the true well-being of the persons holding them. They may be mistaken in thinking that enhancement technology will make them happier or better off. Likewise, those who reject enhancement on the grounds that we ought to be satisfied with what we have (or have been 'given') may wrongly believe that they are well-off and may be wrong to be satisfied with their current condition. This is the well-being problem; it will be briefly discussed in section 8.5.

The other problem is of course that of autonomy. Elster and others have argued that the main problem with adaptive preferences is that they are non-autonomous. Some authors, however, have doubted this conclusion (e.g. Bruckner 2007). In this section it will be discussed why and under what conditions adaptation of preferences impairs autonomy, why this is problematic and what could be done to remedy this.

8.4.1 How Does Adaptation Impact on Autonomy?

How exactly does our unconscious altering of our preferences as a response to our options affect the autonomy of our practical choices? Elster suggests that the main problem with adaptive preferences is that they are *unconsciously* formed and *unintentional* and therefore non-autonomous. According to some authors Elster offers and insufficient or even mistaken explanation of the autonomy problem. Luc Bovens has argued that intentionality is not the point; intentionality according to him is not "a promising candidate as a necessary condition on rational preferences" (Bovens 1992, 67). Bovens offers an alternative way to determine

whether a preference is rational or not. According to him, a preference is rational if it is an all-things considered weighing of the reasons one has to do or want something; this includes the weighing of conflicting desires one may have in light of what one finds important.

I agree with Bovens that intentionality is not the problematic factor. Many of our preferences are not formed intentionally and this does not necessarily mean that they are not rational or autonomous, or worthy of pursuit. As Bovens rightly notes, Elster himself acknowledges that some preferences are autonomous although not intentionally acquired, for example by socialisation, habituation or learning. In this light, it is not even very clear whether the Fox's initial preference for the grapes was formed intentionally.

Likewise, Elster's suggestion that the problem with adaptive preferences is that these preferences have been formed unconsciously, i.e. the subject himself is unaware that his preferences have been influenced by the set of options available to him, does not seem satisfactory. We do not always consciously form or choose our preferences, just like we do not form them intentionally. Many of our preferences are formed unconsciously; they are acquired throughout our lives, through upbringing, habituation and the like. However, I think that as long as one does, upon reflection, acknowledge them as one's own, and as long as they are line with one's goals, plans and values, we can still consider such preferences to be autonomous.

Autonomy and Authenticity

While Bovens' account puts much emphasis on the rationality of preferences, I believe that in discussing the autonomy of preferences in light of the practical choices we make in life, it is also important to focus on the authenticity of these preferences and choices. That is, on the degree to which our preferences and choices can truly be called 'our own'. One of the most important reasons to value autonomy, is that autonomous choices reflect who the person is, and link her choices and actions to her unique individuality. We respect autonomy because we respect persons as beings who are capable of leading their own lives according to their own values and goals (Schermer 2002). Therefore, I propose to understand autonomy as a broader concept than mere rationality and/or intentionality. Being autonomous means being a self-governing person: "being one's own person, directed by considerations, desires, conditions and characteristics that are not simply imposed externally on one, but are part of what can somehow be considered one's authentic self" (Christman and Anderson 2005, 3). Autonomy in this sense is also an ideal that may never be fully reached; but the essential point is that in order to be autonomous in this sense, the desires and preferences that direct ones choices and actions are in the relevant way one's own.

This account of autonomy I use here roughly follows the lines laid out by Frankfurt and Dworkin's theories of autonomy and by consecutive adaptations and refinements made by, among others, John Christman and David DeGrazia (see Christman and Anderson 2005, Taylor 2005, and DeGrazia 2005). In these

theories, the notion of authenticity is central. Being autonomous means leading one's life in accordance with one's own choices, that is, choices based on one's authentic values and goals. On this account an autonomous preference is a preference that conforms to two conditions. First, the subject himself identifies with or endorses the preferences on a higher level of reflection; and second, the process of preference formation, as well as that of identification, must be free of subverting influences. This is the condition of procedural independence.

Procedural Independence and Reflective Endorsement

Colburn (2011) has built on a similar notion of autonomy to argue that Elster's explanation of why adaptive preferences are irrational and non-autonomous is insufficient and to come up with an alternative account. According to Colburn, "autonomy consists in deciding for oneself what is valuable, and living one's life in accordance with that decision" (Colburn 2011, 61). In line with the criterion of procedural independence he argues that what is problematic about Sour Grapes preferences is that they are formed by covert influence. The formation of desires, preferences or other commitments is covertly influenced "when the explanation for those commitments is something that is necessarily hidden from them, in the sense that it would *not* be the explanation for their commitment if it weren't hidden" (Colburn 2011, 64). Thus, covertly influenced preferences are not independent; they crucially depend upon the influence that shaped them. Moreover, this influence is *necessarily* hidden from the subject herself, because if she would be aware of it, it would not be effective in causing the preference.

However, as Bruckner (2011) rightly suggests: the fact that the formation of a preference was not in this sense procedurally independent, does not necessarily undermine autonomy. If reflective endorsement is taken to be the primary criterion for autonomy of preferences, then a person could well come to reflectively endorse a preference that has initially been formed under the influence of something he is not aware of. If the process of preference formation is revealed to him, that is, if he is made aware of the causal explanation that exists for his preference, he can reflect on that and decide whether or not this knowledge is going to change his endorsement of the preference. So, as Bruckner tentatively suggests, the notion of reflective endorsement should include both the endorsement of the preference itself, and the endorsement (or at least non-repudiation) of the process of preference formation. This is very similar to what DeGrazia has argued for in a somewhat different context; he states that the process of preference formation as well as that of identification should not result from "influences that [the person] would, on careful reflection, consider alienating" (DeGrazia 2005, 102).

Following these lines of thought, I propose that in order to determine whether the condition of procedural independence is fulfilled we must determine whether the genesis of the preference and the endorsement of the preference have been free of subverting influences. An influence is subverting if the subject himself

considers it to be so on reflection.[4] This is in line with an inclusive notion of reflective endorsement, i.e. a notion that includes both the preference itself and the process of preference formation. While endorsement could be taken to refer to an emotion or pro-attitude, and be taken to mean something like 'agreeing with' or 'having a positive attitude towards', most authors stress the importance of reflection. According to Bruckner, for example, "the notion of reflective endorsement in play here is not mere endorsement through a second-order preference, i.e., a preference for the preference one has. Rather, the endorsement is an all-in judgment" (Bruckner 2009, 317). Likewise, DeGrazia uses the notion of reflective identification, which also refers to identification with or endorsement of a preference on a higher level of reflection, taking relevant information into account. Reflective endorsement means one holds an all-things-considered positive judgement about one's preferences. In the end, this means (as Bruckner has argued) that what matters is not so much the way in which preferences were formed (their causal origin) but whether or not the person holding them has justificatory reasons to do so, based on his values. Or, as DeGrazia might put it: if there is a reflective harmony in one's system of preferences and values (DeGrazia 2005, 104).

This implies that by reflectively endorsing a preference – including the process of preference formation – one makes it one's own, regardless of how it was formed in the first place. As Bruckner says: "For an adaptive preference to be rational, it need not be autonomously acquired, but we could say it needs to be autonomously retained" (Bruckner 2009, 319). Moreover, I think it is important that identification and reflective endorsement are actual, rather than hypothetical, especially when important preferences are concerned, preferences that guide large parts of one's life or important life-decisions. This does not necessarily mean actual reflective endorsement is needed every time a preference moves one to action, but important preferences that make up one's practical identify need to be reflectively endorsed at some point in one's life in order to be autonomous.

To conclude: the criterion that offers a way to distinguish between problematic, non-autonomous adaptive preferences on the one hand and autonomous character shaping on the other is that of actual inclusive reflective endorsement.

8.4.2 Autonomy of Preferences Regarding Enhancement

Adaptive preferences are not autonomous when they are not reflectively endorsed, because this implies that they are not procedurally independent; their formation depends on the subverting influence of one's circumstances and available options. However, when a person is aware of the influence of his circumstances or option

[4] In his article, Colburn (2011, 65) also suggests that in order to be covert, the influence must not only be hidden in the sense of being unconscious, but must also be such that she would not accept it as a reason for the formation of that preference at all if she was aware of it. This is roughly in line with the proposal made here.

range on his preference formation, or is made aware of them, and then reflects on this, the adaptive preference can either be rejected or reflectively endorsed. In the latter case, it becomes autonomous. We could say that reflective endorsement turns an adaptive preference into a Character Planning-type of preferences. The act of endorsement in this case is very similar to an act of character planning.

This implies that although some preferences contra enhancement as well as some preferences pro-enhancement can be adaptive preferences, as I have argued, and can thus be non-autonomous, they need not *stay* that way. While unconscious adaptation to one's circumstances or (perceived) feasible options is problematic from the point of view of autonomy, when this adaptation is made conscious and the resulting preferences are still endorsed after self-reflection, they become autonomous. By making people aware of the process of preference formation, by making them aware of the way in which unconscious processes may shape their preferences, we offer them the opportunity to reflect on this and to choose whether to endorse or reject this. In that way, their autonomy is supported.

Preferences against Significant Life Extension

I have argued above that some preferences against significant life-extension are indeed adaptive-formed due to the belief that life-extension is not a feasible option, while the person holding such preferences is not aware how the limitations of our present human condition have shaped these preference. However, if one is aware of that, or is made aware of that, e.g. by being informed about the research being done to enhance longevity and the prospects of finding some technology that will indeed enable longer lives, one can reconsider ones preferences. One may, perhaps for the first time, become aware of the fact that we need not always limit our preferences and ambitions to what is actually within our reach, but may legitimately hold ambitions to expand our option set. One may become aware of the fact that this is actually something we do all the time: if we did not have any ambitions beyond our immediate feasible option set, progress and development would be seriously curtailed. One may come to recognize that the only reason one rejected the idea of life-extension in the first place, was because one believed it to be out of reach; and one may come to find that this is not a very good reason to be against the whole project of research and development of significant lifespan extending technologies, but on the contrary, a reason to be in favour of it. So if, after reconsidering one's initial preference against life extension and taking into this consideration the way in which this preference was formed (i.e. under the influence of one's feasible option set) one changes ones preference, this new preference is autonomous.

Likewise, if after such reflection one still endorses the preference against life-extension, this is no longer a non-autonomous adaptive preference; it has become an autonomous conviction. The rationale behind this conviction could take either of two forms. First, it could be an example of character planning: given that the present options do not realistically allow significant life extension, it may be rational or even wise not to form a direct preference for it, but rather to accept that

one will have to do with the 80 or so years that can realistically be expected. This does not exclude the possibility, however, that one can at the same time form an autonomous preference to support the development of technologies for strong lifespan extension, as part of a collective effort to fulfil the desire for longer lives.

Second, it could be that upon reflection one finds additional justificatory reasons to reject technologies for significant life extension; this could imply either a rejection of the use of such technologies for oneself, or a rejection of the whole project of significant life span extension. For example, as many of the citizens in the studies of Partridge and Underwood mentioned, extreme life span extension may cause problems for future generations, like overpopulation, and they may increase unjust inequalities between those who have access to them and those who have not (Underwood et al. 2009, Partridge et al. 2009). These may be good reasons for individuals to *not* form a preference for the development of techniques to significantly increase the human life span. These may also be good reasons for an individual not to prefer extreme life extension for himself – an all-things-considered, reflectively endorsed and therefore autonomous whish *not* to use such techniques if they were available. Moreover, additional personal reasons could also include religious convictions the person feels strongly about. To conclude: regardless of how the initial rejection of life extension came about – whether it was an unreflected adaptation to the status quo or not – in the end, after reflection, the preference against life span extension and in favour of the status quo could very well be autonomous.

Preferences in Favour of Significant Life Extension

In section 8.3.2 I suggested that some preferences in favour of significant life extension could be a special kind of adaptive preferences, namely upwardly adapted or Fisherman's Wife preferences. If preferences to use extreme life-extension technologies (once they become available), or to support or contribute to the development of such technologies are merely based on the fact that they are potentially available, these preferences are not autonomous. If the only reason to want something is because of the fact that you can have it, this preference is not very autonomous; it does not reflect any further reasons based on values, goals or other motivational components of the person himself.

It seems unlikely, however, that anyone would have a preference for life extension technology just because it is – or promises to be – available. Significant life extension is not like a new gadget that one may come to desire just because it is there; and even such a gadget must have some additional appeal besides its availability. There is already something intrinsically desirable about life extension that triggers the preference for life extending technology once this is presented as a potentially available option. It is not just the fact that it may be available that makes it desirable for people – many people hold a desire for longer lives even though they are well aware that this is not (at the moment) a realistic option. They simply argue that if life is good and valuable, more life must be even better.

Opening up of options that were previously thought to be unavailable, or even unimagined, may spark preferences for those new options. But that does hardly ever mean that the availability is the *only* reason to have those preferences – they will have to appeal to some element in the person's set of values, goals and desires to appear as attractive for this person. This attractiveness may lie in some property of the newly available option itself, or in the fact that it is considered to be something new, exciting, fashionable or distinctive in some way. The appeal of the latter type of qualities is what is consciously manipulated and exploited by merchandising and advertising of consumer goods. The preferences for certain enhancement technologies may also depend for an important part on this kind of appeal. Cosmetic enhancements, for example are very fashion-sensitive. And some scientific and biomedical researchers in the various fields of the Converging Technologies[5] appear to be driven mainly by the exciting, new and pioneering nature of their creations.

Nevertheless, when life extension and extreme longevity are considered, it is reasonable to suppose that most people who value life will feel the appeal of the option of life extension. Just like it is not unreasonable to suppose that most people – just like the Fisherman's wife – would, at first sight, like to have a bigger house or even a villa, or a castle, if it was made available to them. Even though this actual conscious desire may be triggered by the offer, or the opening up of the option, it corresponds to some values, goals or desires the person already holds. This does not mean, however, that on reflection the person would endorse this desire. The desire may not have been formed merely in reaction to the feasible option set, it is triggered by that option set. It is not a reflectively endorsed, all-things-considered preference. After careful consideration one may come to find reasons against life-span extension, like the ones mentioned earlier and one may realise that the first intuitive appeal of life extension is not warranted. Alternatively, further conscious deliberation could also result in reflective endorsement of the desire for life extension. Upon reflection, one may thus either change one's preferences, or come to re-affirm them. In both cases, however, these preferences are more autonomous than the initial desire which was triggered by the perception that life-span extending technologies are within our reach.

The lesson that can be drawn from the discussion in this section is that unconscious adaptation of preferences either for or against enhancement to the (perceived) available options renders these preferences less than fully autonomous. Through reflective endorsement, these preferences can become autonomous. Alternatively, preferences may change after reflection on their formation, and in that case the new preferences that are reflectively endorsed, can be considered to be autonomous.

One can thus autonomously desire enhancements like extreme lifespan extension (even if these are still promises and not effectively available options)

[5] Converging technologies is a term referring to the increasing interaction and mutual reinforcement of Nanotechnology, Biotechnology, Information and Communication Technology and Neuroscience.

but one can also be autonomously satisfied with a non-enhanced life with the presently available lifespan. Both positions deserve respect in the debate. However, to the extent that preferences regarding enhancement – for example preferences regarding life-extension – are non-reflectively formed in direct response to their perceived availability or lack of availability, they lack autonomy and rationality, and need not be taken too seriously. Open discussion on the mechanisms of preference formation, as well as of the pros and cons of significant life span extension – or other enhancements – is necessary to turn such unexamined desires into truly autonomous preferences.

8.5 Adaptation, Well-Being and Enhancement

In this final section I will briefly and tentatively turn to the other problem that adaptive preferences pose, next to the autonomy problem: the problem of well-being. As introduced in section 8.2.2, the noted differences between peoples' subjective appreciation of their quality of life as contrasted to the objective or third person evaluations of this quality has sparked interest in the notions of adaptation and adaptive preferences in the fields of development studies and healthcare. The problem is that through adaptation to miserable, deprived, suppressed or disabled circumstances individuals fail to recognize how badly off they actually are, and fail to desire the things they ought to desire. They are satisfied with situations, preferences and circumstances they should not be satisfied with. Serene Khader has argued that the problem with adaptive preferences in the context of human development cannot be captured by reference to the lack of autonomy of such preferences. According to her, even preferences that are autonomous on most accounts of autonomy can suffer from the problem that they do not reflect a person's true well-being as most of us would see it. Therefore, she contends that understanding the problem of well-being and adaptive preferences "requires a theory of the good" (Khader 2009, 169).

Indeed, the problem of well-being in relation to adaptation of preferences as it poses itself in the field of human development seems to require a certain idea of the good, or of human flourishing. One of the main tasks capability theorists have set for themselves is precisely to determine what the objective conditions for human flourishing are. This is essentially a normative task. The set of capabilities that everyone ought to have, according to capability theorists, refers to a set of options and choices that every person should have or should have a right to have – one should not settle for less, one should not adapt to less than this minimal set. Those who have a smaller feasible option set but nevertheless proclaim to be happy, suffer from adaptive preferences, according to this view. Circumstances that fail to provide the minimal capabilities and that produce such adaptive preferences are considered unjust. The capability approach thus formulates a minimal set of options that people have a right to desire and that governments have a duty to provide.

Proponents of enhancement can[6] make a very similar normative claim: that we as human beings ought to have a certain opportunity range available to us and that we should not let biological factors or 'natural' forces prevent us from having that. While the capability approach directs its attention to socio-political limits to our capabilities or options, proponents of enhancement put the natural limits up for discussion. Just like socio-political limitations natural limits can be undesirable and need not be accepted, in their view. Biotechnology has brought so much of our biology – of our 'nature'– within the range of human intervention powers that it can no longer be considered to fall outside the realm of ethics and justice. Human beings have so much power to intervene in human biology that this cannot be accepted as a mere given anymore – human biology for a large part has come to lie within the range of human responsibility. Although not undisputed, John Harris for example has therefore claimed that it is a moral duty to try to enhance ourselves by biotechnological means (Harris 2007). More specifically directed at the idea of life extension: pro life-extension activists claim that as a matter of justice, research into anti-aging should be stimulated by governments, since "aging kills 100.000 a day and it is not inevitable!" (http://longevityalliance.org/the-brussels-summit-of-longevity-activists)

While the capability approach can be understood to define the contours of human flourishing and especially to define the *minimal* requirements for the possibility of leading good human life, the pro-enhancement stance presents human flourishing as open ended, flowing over into post-human flourishing. There is in principle no upper limit to what we may aspire to. Especially in the writings of transhumanists and proponents of radical enhancement there is no natural or other boundary to the options we may form desires for: immortality; enhancing our intelligence billionfold; having new sensibilities and modalities like infrared vision or telepathic sense; these are all options we could or even should form desires for, according to transhumanists.[7] This opens up an endless space of potential desires and therefore endless opportunity for upward preference adaptation. While the capability approach aims to set a minimal standard for the options we need for human flourishing, the proponents of human enhancement seem to stress that there is no maximum to flourishing or to the options we should pursue. This seems to imply that according to the transhumanist point of view simple mortal human beings must be dissatisfied and unhappy with who they are and what they have. This situation brings back to mind the predicament of the Fisherman's wife, who always wanted more and was never really happy or satisfied with what she got. In turn this raises the question how well-off someone in such a predicament really is, because however well-off someone in such a predicament objectively is, in light of ever greater possibilities this will never be enough to feel complete subjective well-being. This seems to be a caveat for

[6] I mean to show that this is a possible way of thinking, found by some, but certainly not all, transhumanists or other strong proponents of enhancement.

[7] For more examples see www.nickbostrom.com/ethics/values.html

theories that state the human condition can and should be improved indefinitely. We could say that the tension between subjective and objective evaluations of well-being which is a core-issue in the capability approach (as well as in many discussions about quality of life in healthcare) is turned upside down in the enhancement discussion. Instead of warning against the situation of the Happy Slave and stressing the inadequacy of a merely subjective evaluation of well-being, as the capability approach does, in the enhancement debate we should perhaps be more aware of the predicament of the Fisherman's wife, and re-evaluate the importance of subjective well-being in situations of sub-maximal objective well-being. Here, we may draw some lessons from the disability movement, where the disability paradox has sparked similar discussions (Goering 2008, Barnes 2009).

In discussing these issues, it is inevitable that the question of how the concept of 'true well-being' is linked to reflective endorsement of preferences and hence to autonomy will come up. According to the authentic happiness account of Sumner well-being consists in the informed and autonomous judgments of people about their own life-satisfaction and fulfilment (Sumner 1996). As mentioned above, however, Khader argues that such an account will not do, at least not in the context of human development. A full discussion of the nature of well-being and its precise relation to autonomy goes beyond the focus of the present volume. However, in order to fully understand the role preference adaptation for well-being in the context of the enhancement debate, such a discussion would need to be conducted.

8.6 Conclusion

In this chapter I have explored the role that adaptive preferences play in the debate on human enhancement, especially with regard to lifespan extension and prolongevity. Since the notion of adaptive preferences is conceptualised somewhat differently in different theoretical fields, it draws our attention to two distinct sets of problems. First, the problem of the autonomy of our preferences and second the problem of the evaluation of our well-being. I have argued that individual preferences for as well as individual preferences against certain forms of enhancement can be adaptive, in the sense that they are formed unreflectively in response to one's perceived options. Preferences in favour of life-extension are not strictly speaking adaptive preferences in that sense because they are triggered, rather than formed, by their perceived or promised availability. Nevertheless, in all cases where preferences regarding lifespan extension, or other enhancements, are formed unconsciously and are not reflectively endorsed they are not fully autonomous. Upon reflection on the formation and content of one's enhancement-related preferences, individuals can either come to endorse or to change their initial desires, thus making them more 'their own' and therefore more autonomous. Since autonomy is an important value that is closely linked to respect for persons and respect for their self-governing choices, we should aim at the

highest possible levels of autonomy when it comes to preferences and choices regarding lifespan extension and other enhancements. It is useful to consider the phenomenon of adaptive preferences in the context of the enhancement debate because this stimulates critical reflection on the process of enhancement-regarding preferences and can therefore help to improve the autonomy of such preferences.

Secondly, the notion of adaptive preferences also offers a new way of looking at what I take to be the ultimate goal of enhancement: to improve human well-being. I have argued that some preferences in favour of enhancement may be examples of upward adaptation and questioned how happy or well-off the quest for ever more options actually makes us. While the capability approach to adaptive preferences warns against taking subjective evaluations of well-being too seriously, the lesson for the enhancement debate may be that we should consider the effects that a limitless view of human flourishing may have on our subjective sense of well-being and life satisfaction.

References

Albrecht, G.L., Devlieger, P.J.: The Disability Paradox: High Quality of Life Against All Odds. Soc. Sci. Med. 48, 977–988 (1999)

Barnes, E.: Disability and Adaptive Preference. Philos Perspect 23, 1–22 (2009)

Bostrom, N., Ord, T.: The Reversal Test: Eliminating Status Quo Bias in Applied Ethics. Ethics 116, 656–679 (2006)

Bostrom, N.: Human Genetic Enhancements: A Transhumanist Perspective. J. Value Inq. 37, 493–506 (2003)

Bovens, L.: Sour Grapes and Character Planning. J. Philos 89, 57–78 (1992)

Bruckner, D.W.: In Defence of Adaptive Preferences. Philos Stud. 142, 307–324 (2009)

Bruckner, D.W.: Colburn on Covert Influence. Utilitas 23, 451–457 (2011)

Buchanan, A.: Beyond Humanity? Oxford University Press, Oxford (2011)

Christman, J., Anderson, J. (eds.): Autonomy and the Challenges to Liberalism. Cambridge University Press, Cambridge (2005)

Clark, D.A. (ed.): Adaptation, Poverty and Development: The Dynamics of Subjective Well-Being. Palgrave Macmillan, New York (2012)

Colburn, B.: Autonomy and Adaptive Preferences. Utilitas 23, 52–71 (2011)

DeGrazia, D.: Human Identity and Bioethics. Cambridge University Press, Cambridge (2005)

Elster, J.: Sour Grapes: Studies in the Subversion of Rationality. Cambridge University Press, Cambrigde (1983)

Gilbert, M.: Collective Preferences, Obligations, and Rational Choice. Econ. Philos 17, 109–119 (2001)

Goering, S.: 'You Say You're Happy, But...': Contested Quality of Life Judgments in Bioethics and Disability Studies. J. Bioethical Inq. 5, 125–135 (2008)

Harris, J.: Enhancing Evolution. Princeton University Press, Princeton (2007)

Khader, S.J.: Adaptive Preferences and Procedural Autonomy. J. Hum. Dev. Capabil 10, 169–187 (2009)

Nussbaum, M.: Women and Human Development. Cambridge University Press, Cambridge (2000)

Ong, A.D., Bergeman, C.S., Boker, S.M.: Resilience Comes of Age: Defining Features in Later Adulthood. J. Pers 77, 1777–1804 (2009)

Post, S.G., Binstock, R.H. (eds.): The Fountain of Youth: Cultural, Scientific, and Ethical Perspectives on a Biomedical Goal. Oxford University Press, Oxford (2004)

Ryff, C.D., Friedman, E.M., Morozink, J.A., Tsenkova, T.: Psychological Resilience in Adulthood and Later Life: Implications for Health. Annu. Rev. Gerontol. Geriatr. 32, 73–92 (2012)

Schermer, M.: The Different Faces of Autonomy. Kluwer Publishers, Dordrecht (2002)

Schermer, M., Pinxten, W. (eds.): Ethics, Health Policy and (Anti-)Aging: Mixed Blessings. Springer, Heidelberg (2013)

Sumner, L.W.: Welfare, Happiness and Ethics. Oxford University Press, Oxford (1996)

Taylor, J.S. (ed.): Personal Autonomy: New Essays on Personal Autonomy and its Role in Contemporary Moral Philosophy. Cambridge University Press, Cambridge (2005)

Teschl, M., Comim, F.: Adaptive Preferences and Capabilities: Some Preliminary Conceptual Explorations. Rev. Soc. Econ. 63, 229–247 (2005)

Partridge, B., Underwood, M., Lucke, J., Bartlett, H., Hall, W.: Ethical Concerns in the Community about Technologies to Extend Human Life Span. Amer. J. Bioethics 9, 68–76 (2009)

Underwood, M., Bartlett, H.P., Partridge, B., Lucke, J., Hall, W.: Community Perceptions on the Significant Extension of Life: An Exploratory Study Among Urban Adults in Brisbane, Australia. Soc. Sci. Med. 68, 496–503 (2009)

Chapter 9
Self-Deception, Adaptive Preferences, and Autonomy

James Stacey Taylor

9.1 Introduction

In Neville Shute's (2010) novel *On the Beach*, set in Australia after World War III, one of the main characters, Commander Towers, buys his children gifts and imagines them getting older–even though he is American and knows that all human and animal life in the northern hemisphere is almost certainly extinct. As presented in the novel Towers' actions were clearly the result of self-deception. This is not surprising, for many people deceive themselves in the face of death or serious illness; the phenomenon of patients who have been informed of their terminal illness failing to account for this in their plans is well known to healthcare professionals.[1] On the face of it, this phenomenon accords well with the widely held belief that a person who engages in self-deception will suffer from diminished autonomy with respect to her decisions that result from this. After all, it is widely held that when a person is deceived into holding certain preferences or making certain decisions she is not autonomous with respect to them, and

James Stacey Taylor
Department of Philosophy
The College of New Jersey
Ewing, NJ
The United States
e-mail: jtaylor@tcnj.edu

[1] See, for example, Hamlyn (1985). But it is not only patients whose decisions might be the result of self-deception; healthcare professionals might also engage in self-deception when called upon to make life or death decisions (such as the decision as to whether or not refuse to admit a patient to hospital), as a result of being unwilling to acknowledge that the refusal to provide life-prolonging treatment could be the moral equivalent of intentionally causing death. See here Savers and Perera (2002).

J. Räikkä and J. Varelius (Eds.): *Adaptation and Autonomy*, SAPERE 10, pp. 137–148.
DOI: 10.1007/978-3-642-38376-2_9 © Springer-Verlag Berlin Heidelberg 2013

self-deception appears merely to be a subset of the ways in which a person can be deceived successfully.[2] And just as it is widely held that successfully subjecting oneself to self-deception will compromise one's autonomy, so too is it widely held that persons will suffer from diminished autonomy with respect to their adaptive preferences–those preferences that persons form as a result of their downgrading of options that they believe are inaccessible to them.[3] Thus, for example, an elderly person who had always valued the independence of living in her own home might come to prefer the comfort and dependency of an assisted living facility after suffering a debilitating stroke that precluded her from safely living as she did before. Given that she had this preference only as a result of her perception of the contraction of her feasible options it was adaptive–one that she possessed to reconcile herself to her perceived more constrained circumstances.

Given their association with situations that are perceived to be harmful by the person whose motivational set is affected by them cases of self-deception and adaptive preference are not uncommon in healthcare settings, especially those that involve end of life care. But that this is so raises an important question: If persons in healthcare settings, especially those associated with end of life care, are susceptible to self-deception and adaptive preference formation, and if both of these phenomenon independently compromise autonomy, should a concern for patient autonomy give way to a greater focus on beneficence in contemporary healthcare?

The aim of this essay is to argue that independently of the merit or otherwise of the current focus on respecting patient autonomy this should not be replaced by a focus on beneficence on the grounds that persons in (especially) end of life healthcare settings are susceptible to suffering from compromised autonomy as a result of their tendencies to self-deception and adaptive preference formation.[4]

---

[2] See, for example, Baron (1988, 436). I accepted the view that self-deception could compromise autonomy in Taylor (2009, 73-74). I now believe that I was mistaken there.

[3] This account of what constitutes an adaptive preference is similar to that of Jon Elster, who describes an adaptive preference as one that a person forms unconsciously that downgrades options that are inaccessible to them (Elster 1987, 119). Note, however, that the account of adaptive preference used in this paper is broader than that of Elster's, for it accepts that a person's preferences can be adaptive even if she has consciously taken steps to revise them to fit better with the constraints that she perceives herself to be subject to. While persons who wish to adhere to Elster's narrower conception of adaptive preference will find little of interest in this discussion of adaptive preference and intentional self-deception (since they will reject the claim that the preferences under discussion are genuinely adaptive), the next section will still be of interest to them since it focuses on preferences that persons hold as a result of unintentional self-deception, which accords with Elster's view that adaptive preferences are unconsciously formed.

[4] I have argued that we should reject autonomy as the central value of medical ethics in Taylor (2009, Ch. 10).

This is because, it will be argued, neither self-deception nor adaptive preference formation compromise the autonomy of those who are subject to them. Rather than treating these as separate issues, however, this essay will treat them together, focusing on adaptive preferences that arise as a result of self-deception. Given the conclusion of this paper combining these issues in this way will be advantageous, for if it can be shown that even an adaptive preference that arose through self-deception is one that its possessor can be autonomous with respect to then there will be no bar to a person being autonomous with respect to either her (non-adaptive) preferences that she holds as a result of self-deception, or her adaptive preferences.

Prior to arguing that persons can be autonomous with respect to even an adaptive preference that was formed through self-deception it must be admitted that the precise nature of self-deception is a matter of controversy.[5] Some persons hold that self-deception occurs intentionally; others deny this, and claim that it occurs non-intentionally. To avoid pinning the argument of this paper to either of these controversial positions it will be argued that a person could be autonomous with respect to his adaptive preferences that have their origins in self-deception whether the self-deceptive processes that led him to have them were intentional or non-intentional.

9.2 Intentional Self-Deception and Adaptive Preferences

To this end, let us first consider first intentional self-deception. On one prominent approach to analyzing self-deception this phenomena takes place across time. A self-deceiver believes that X, and wants to come to believe that ~X. To achieve this cognitive change she sets out to deceive herself into believing ~X, and as she does so either forgets her original deceptive intent and so comes to believe that ~X, or else remembers it and yet come to believe that she has arrived at a true belief (~X) through an aberrant process.[6] Take, for example, Bully Stryver, the egotistical lawyer from Charles Dickens' (1998) *A Tale of Two Cities*. Let us assume that Stryver decided to deceive himself into believing that he did not want to marry Lucie when she turned down his proposal of marriage.[7] To this end he concentrated on her bad qualities (real or imagined), listened intently to his friends who told him that she wasn't right for him anyway, and, as a result, came to believe that he was misguided to have wanted to marry her and formed the adaptive preference not to. It is clear that there is no bar to Stryver being

[5] Indeed, so is its very existence; see Paluch (1967).

[6] This is discussed in Sorenson (1985).

[7] The assumption that Stryver acts in this way is made to avoid begging the question against non-intentional accounts of self-deception. The example of Bully Stryver is chosen for a reason that will become clear below.

autonomous with respect either to his decision to deceive himself or to his subsequent self-deceptive course of action. But the question of whether or not he is autonomous with respect to his self-deceptively generated adaptive preference is separate from this, for a person might autonomously bring herself into a nonautonomous state. And, at least at first sight, there is reason to believe that Stryker would *not* be autonomous with respect to his self-deceptively engendered adaptive preference not to marry Lucie. On Christman's (1993, 288) historical account of autonomy, for example, an agent P is autonomous with respect to a desire D at time t if

i. P did not resist the development of D (prior to t) when attending to this process of development, or P would not have resisted that development had P attended to the process.
ii. The lack of resistance to the development of D (prior to t) did not take place (or would not have) under the influence of factors that inhibit self-reflection;
iii. The reflection involved in condition I is (minimally) rational and involves no self-deception; and
iv. The agent is minimally rational with respect to D at t (where minimal rationality demands that the agent experience no manifest conflicts of desires or beliefs that significantly affect the agent's behavior and that are not subsumed under some otherwise rational plan of action.

Clearly, for Christman, Stryver would not be autonomous with respect to his adaptive preference not to marry Lucie, for he would fail to meet condition (iii) to be autonomous with respect to it. But that Christman would hold that Stryver failed to be autonomous with respect to his self-deceptively formed adaptive preference not to marry Lucie does not definitively show that he lacked autonomy here, for Christman merely stipulates that persons are not autonomous with respect to preferences formed through self-deception, rather than offering reasons to accept this view. To be sure, and in fairness to Christman, it is not implausible to hold that a preference that is held as a result of self-deception is not one that its possessor is autonomous with respect to.[8] After all, if Cordelia deceives Charles into possessing a preference P he would not be autonomous with respect to it, for it would be she, and not he, who directed him to have it. But cases of interpersonal deception differ crucially from cases of self-deception. When one person successfully deceives another into possessing a particular preference or making a particular decision it is she, and not he, who directs him to possess it or perform it. However, when a person deceives herself her possession of the preference or performance of the action in question is still directed by *her* self, and *not* by

[8] Indeed, I accepted this view uncritically in Taylor (2009).

another–and so without further argument there is no reason to believe that she is not autonomous, self-directed, with respect to it.[9]

Perhaps, though, one might believe that a person who is *self*-deceived into possessing a preference P is not autonomous with respect to it because in pursuing its object she will be pursuing something that she has greater reason not to pursue, and she would recognize this were she not self-deceived. But it is not clear why a person cannot be autonomous with respect to a preference P or a decision to perform action A when she has more reason to prefer Q to P, or to decide to perform B rather than A. While there is a clear link between autonomous action and rational action in the Kantian tradition of moral autonomy, no such link holds when "autonomy" is understood (as it is here) in the sense of personal autonomy, where a person is autonomous with respect to her decisions, preferences, desires, and actions if they can be genuinely held to be hers.[10] As such, even if it is the case that a person has more reason to prefer Q to P, or to perform act B rather than act A, this is no *a priori* bar to her being fully autonomous with respect to P or A. Moreover, it is plausible to hold that adaptive preferences that are the result of self-deception *are* those that a person has most reason to possess. Bully Stryver's adaptive preference not to marry Lucie, for example, was instrumentally valuable for him insofar as it enabled him to preserve his self-respect (see Richardson 2001, 71-72). Thus, even if one believes that a person is not autonomous with respect to those of her preferences whose objects she has greater reason not to pursue one could still accept that a person could be autonomous with respect to those of her adaptive preferences that she deliberately inculcated in herself through self-deception.

---

[9] One might object at this point that this response overlooks another possibility–that a person is autonomous only when she is directed by her "true" or "authentic" self, and that self-deception consists of her "inauthentic" self deceiving her "authentic" self. (An account of self-deception similar to this has been developed by (Pears (1991).) If this is so, then the adaptive preferences that a person would hold as a result of such intentional self-deception would be ones that she would not be autonomous with respect to, for her possession of them would be the result of her being subject to the deception of her inauthentic self–and if successful such deception would compromise the autonomy of her true self just as much as would successful deception by another agent. There are two responses to this line of objection. First, one might argue against the "true self" view of autonomy that it is based on–see, for example, Friedman (1986). (Note, too, that the account of autonomy that is drawn upon in this paper is not a "true self" view–see Taylor (2009, Chs. 1, 7) Second, one could note that this view of self-deception borders on being incoherent. It draws its plausibility from the combination of the view that the "authentic" self of the person subject to self-deception is deceived by her "inauthentic self", and so her autonomy is compromised in a manner that is analogous to how it would be compromised were she to be successfully subject to interpersonal deception. But for this to be an account of *self*-deception the self that is deceiver must be the same self that is the deceived. Thus, while analogizing self-deception to interpersonal deception in this way does explain how self-deception could compromise autonomy it does so by undermining the claim that the deception at hand is truly self-deception.

[10] Note that this characterization of personal autonomy has been left deliberately vague, to admit of different analyses of what is required for a person's decisions, preferences, desires and actions to be "truly hers".

Thus far, then, there is no reason to believe that a person (like Bully Stryver or Commander Towers) could *not* be autonomous with respect to an adaptive preference that she has brought herself to have through intentional self-deception. But this does not give us reason to believe that we should accept that under some circumstances a person is autonomous with respect to such a preference. To provide this, consider both the first part of the Threshold Condition and the Degree Condition that I developed as part of my analysis of what conditions must be met for a person to be autonomous with respect to her decisions (and hence the pro-attitudes and actions that flow from them):

> Threshold Condition: It is necessary for a person to be autonomous with respect to a decision that she makes that (i) the information on which she based the decision has not been affected by another agent with the end of leading her to make a particular decision, or a decision from a particular class of decisions (Taylor 2009, 7).[11]
>
> Degree Condition: The maximum degree to which a person will be autonomous with respect to a decision that she makes will be determined by the degree to which it is the result of a decision-making procedure that she is satisfied with as being her decision-making procedure for making the type of decision that is in question; to the degree that the genesis of her decision departs from this she will suffer from a diminution in her autonomy with respect to the decision in question (Taylor 2009, 8).[12]

---

[11] Parts (ii) and (iii) of the Threshold Condition–which is a necessary but not a sufficient condition for a person to be autonomous with respect to her decisions–concern the influence of other agents on the information that a person bases her decision on. Since the discussion in this paper focuses on self-deception, these are not relevant to it, and so will not be outlined here. For a discussion and defense of them, see Taylor (2009, 7).

[12] Note that this Condition is neither as strict nor as lax as it might appear. It is not as strict as it might appear for in requiring that a person make her decisions in accord with a decision-making procedure that she is satisfied with all that is required is that she have reflected upon how she makes the type of decision that is question, and be satisfied with this procedure. She thus need not engage in such endorsement of her decision-making procedure every time she makes a decision, but to be autonomous with respect to her decisions she must have endorsed the procedure (in the sense of reflecting on it and being satisfied with it) at some point in the past, and continue to be satisfied with it. To be sure, this might still be held to be an overly-strict requirement, for it would preclude persons from being autonomous with respect to many of their everyday decisions, if these are relatively so trivial that they have not reflected on the procedures by which they made them. Yet this does not imply that persons are thereby alienated from these decisions, for there is nothing in this theory of autonomy that commits it to a binary distinction between autonomy and heteronomy. Indeed, not only is it possible that persons can act or decide agentially (i.e., as non-autonomous, non-alienated agents), but it is plausible that many of the everyday acts and decisions that persons perform or make are of this sort. It is also not as lax as it might appear, for although this is a satisfaction-based condition a person will only count as being satisfied with her decision-making procedures on this Condition if she has reflected upon them; mere acceptance will not suffice.

The final Condition that must be met for a person to be autonomous with respect to her decisions is the Tracing Condition, which concerns those circumstances when a "person decides to use a different decision-making procedure than that which is satisfied with as being her own…" (Taylor 2009, 11). Since this is not directly relevant to typical cases of adaptive preferences that are formed through intentional self-deception it is not discussed here.

Bully Stryver's decision to deceive himself so that he developed adaptive preferences were made on the basis of information that had been affected by another with the aim of leading him to make a particular decision, or a particular type of decisions. And, *ex hypothesi*, his decision was the result of decision-making procedures that he was satisfied with for making the type of decision that this was.[13] Moreover, the decisions that he made on the basis of his self-deceptively inculcated adaptive preferences also satisfied these two Conditions: they were not made on the basis of information that had been affected by another agent with the aim of influencing his decisions, and they were the products of decision-making procedures that he was satisfied with. As such, then, Stryver satisfies the above conditions to be autonomous with respect to the decisions that he made as a result of having adaptive preferences that he had inculcated into himself through self-deception. That a person has and acts upon a self-deceptively engendered adaptive preference, then, does not thereby show that she suffers from any diminution in her autonomy.

9.3 Unintentional Self-Deception and Adaptive Preferences

Stryver could thus be fully autonomous with respect to his adaptive preferences if they were the result of intentional self-deception. But what if they were the result

[13] One might object that he could have had a policy not to make decisions on the basis of information that he had self-deceptively presented to himself, and hence he would fail to satisfy the Degree Condition to be autonomous with respect to decisions that he made as a result of being self-deceived–including those that he made on the basis of the adaptive preferences that he possessed as a result of self-deception. But to object in this way would be to confuse being satisfied with a decision-making *procedure* with being satisfied with the informational inputs *into* that procedure–and it is only the former that the Degree Condition is concerned with. Admittedly, this response has an air of stipulation about it. To make it more palatable, it should be noted that requiring that for a person to be autonomous with respect to her decisions she must be satisfied with the quality of information that they are based on would render persons non-autonomous with respect to their decisions if they desired that they be made on the basis of accurate information (plausibly, a widespread desire) but were not. On the face of it, this might not appear to be an implausible result. But it would implausibly render persons non-autonomous with respect to their decisions even if accurate information was not accessible to them through no fault of their own or of others–and this is implausible. It is not, for example, plausible to hold that Newton was non-autonomous with respect to his scientific work because his decisions were made in ignorance of quantum mechanics, or to hold that nineteenth century physicians wee non-autonomous with respect to their medical work because they acted in ignorance of the germ theory of disease. For further argument against the view that persons must have a robust understanding of their actions and their implications to be autonomous with respect to them–that is, for an argument against the view that the quality of the informational inputs into a person's affect the degree to which she is autonomous with respect to the decisions in which they play a role–see Taylor (2013).

of *non*-intentional self-deception? In such cases the belief that is the product of self-deception and that leads the persons concerned to possess their adaptive preferences is motivated by a conative state whose object is the content of the belief (see, for example, Barnes 1997 and Mele 2001). On this analysis of self-deception Stryver would be motivated by his desire to protect his self-respect to believe that Lucie would make a poor wife. Here, beliefs would count as self-deceptive if their possessors either recognized at some level that they were likely to be false, and so made efforts to resist acknowledging this, or else (and more weakly) that they were uncertain whether they were true or not but adopted a belief that served a positive conative function for them in the face of evidence to the contrary.

As with adaptive preferences that were the result of intentional self-deception it seems plausible to hold that adaptive preferences that are the result of non-intentional self-deception would be preferences that persons would not be autonomous with respect to.[14] After all, the beliefs that they are based upon were formed by a process that, while it served the interests of the person concerned, was neither under her control nor one that she was aware of. Were a person thus to form a preference that was based on a belief generated by such a covert process it would be the result of something that lay beyond the scope of her self-direction–and so it is plausible that she would not be autonomous, would not be self-directed, with respect to it.

Yet while initially plausible, it is important to note that this account of how an adaptive preference that was the result of unintentional self-deception would be one that its possessor is not autonomous with respect to commits its proponents to holding that a person's autonomy with respect to her preferences, decisions, or actions would be compromised if their motivational ancestry included any cognitive or conative mental states that had their origins in processes that she was not aware of. But given that many of a person's beliefs and pro-attitudes arise from processes that she is not aware of this view of what factors serve to compromise a person's autonomy would lead to an implausibly narrow view of which preferences, decisions, and actions a person is autonomous with respect to. Of course, this alone is not a fully satisfactory response to the view that forms the foundation of this account of how an adaptive preference that was the result of unintentional self-deception would be one that its possessor was not autonomous with respect to, for its proponents could simply retort that this just shows that autonomy is a rarer property of persons than was previously thought. To respond properly to this account of how an adaptive preference that was formed through unintentional self-deception is not one that its possessor is autonomous with respect to, then, it will be necessary to criticize its theoretical foundations.

[14] This appears to be the view of Elster, who writes that such preferences are formed "behind the back of the agent concerned" (Elster 1987, 117); see also Zimmerman (2003, 221) and Colburn (2011, 57).

The intuitive force of the view that a person is only autonomous with respect to those of her preferences, decisions, or actions whose motivational ancestry she is aware of springs from the view that to be autonomous with respect to a preference, decision, or action a person must have at least endorsed (if not actually directed herself to have) her possession of those of her mental states that led her to possess it. Were this not to be the case, the proponents of this view continue, then she would be passive with respect to its genesis; it would not be something that she directed herself to possess (and hence something that she is autonomous with respect to), but merely something that was instilled into her (and hence something that she is heteronomous with respect to).[15] But such a strongly voluntaristic view of autonomy faces a serious theoretical difficulty. The endorsement of a mental state S that one possesses must be grounded in other mental states that one has–one's beliefs about S, as well as one's conative states concerning S and one's beliefs about these. On a strongly voluntaristic view of autonomy, then, for one's endorsement to be an act that one was autonomous with respect to one must have endorsed these mental states–as well as the processes that led to them–in the past. But this would, again, require a series of *prior* autonomy-conferring endorsements–as would these endorsements also. Since there could not have been an infinite regression of such endorsements (persons are, after all, finite!), and since on this strongly voluntaristic view an unendorsed mental state would undermine the autonomy of the person who possessed it with respect to those of her preferences in whose ancestry it featured, there must be some mental states that their possessors simply picked to endorse, without any basis for this picking. But, if this is so, then it is unclear as to why one should place any value on autonomy at all. After all, if on this view any autonomously-held preference must have its origins in a mental state that was simply picked, why should that preference with its arbitrary origin possess any more normative force (either for its possessor or for any third party) than any other?[16]

Given this serious difficulty faced by the strongly voluntaristic view of autonomy that undergirds the above view of why a person is not autonomous with respect to an adaptive preference that originated from unintentional self-deception, it would be wise to turn elsewhere for an account of autonomy. Clearly, in the light of this discussion satisfactory account of autonomy must be one that

[15] The association of passivity with heteronomy was introduced by Frankfurt (1999, 133). This voluntaristic approach to autonomy seems to be that adopted by Ben Colburn, which is why he holds that persons are not autonomous with respect to their adaptive preferences that arise through processes that they are not aware of; he holds that "Autonomy consists in deciding for oneself what is valuable, and living one's life in accordance with that decision" (Colburn 2011, 61).

[16] A more thorough and extended discussion of these issues, albeit outside the context of self-deception and adaptive preferences, can be found in Taylor (2009, Ch. 7).

recognizes that autonomous self-direction does not require that a person also direct (i.e., form) the self that is the font of her motivations. Rather, a person will be autonomous with respect to her preferences if they arise without undue influence from another. This is reflected in part (i) of the Threshold Condition, above. But as it stands this is too restrictive, for it would preclude a person being autonomous with respect to a decision that she had made on the basis of the advice of another, after she had weighed this and found it sound. It would also preclude a person from being autonomous with respect to her decision if she made it based on information that was presented to her by someone who was attempting to manipulate her when this manipulation was unsuccessful, and so her decision was not that which her would-be manipulator intended. Given the overly restrictive nature of (i), then, it should be supplemented by:

> (ii) if the information on which a person makes her decision has been affected by another agent with the end of leading her to make a particular decision, or a decision from a particular class of decisions, she is aware of the way in which it has been so affected (Taylor 2009, 7).

And

> (iii) if the information on which a person bases her decision has been affected by another agent with the end of leading her to make a particular decision, or a decision from a particular class of decisions, and if she is not aware of the way in which the information on which she is basing her decision has been so affected, then she did not make the decision that the agent who was affecting the information she had access to with the intent of leading her to make a particular decision intended her to make (Taylor 2009, 7).

With the above rejection of the strongly voluntaristic view of autonomy in hand, together with this full account of the Threshold Condition, it is clear that (when this condition is combined with the Degree Condition and Tracing Condition, above, to provide a full account of practical personal autonomy) there is no bar to a person being autonomous with respect to her adaptive preferences that arose as a result of non-intentional self-deception. Since the information on which she bases her decision has not been affected by another agent with the end of leading her to make a certain decision, or a decision of a certain type, she meets the Threshold Condition to be autonomous with respect to those of her decisions that she makes on the basis of her adaptive preferences, and hence autonomous with respect to the preferences themselves. Since the mere fact that her preferences are adaptive and arose through non-intentional self-deception does not preclude her from making her decisions on the basis of "a decision-making procedure that she is satisfied with as being her decision-making procedure for making the type of decision that is in question" she can meet the Degree Condition for her to be autonomous with

respect to her decisions made on the basis of them–and hence autonomous with respect to them. And, similarly, there is no bar to her meeting the Tracing Condition.[17]

9.4 Conclusion

It was noted at the start of this paper that adaptive preferences–including adaptive preferences that are generated through self-deception–are not uncommon in healthcare settings, especially those that involve end of life care. Since it is frequently believed that persons fail to be autonomous with respect to their adaptive preferences–and hence with respect to the decisions that they make on the basis of them–one might conclude from their prevalence within healthcare settings that healthcare settings are fora for compromised autonomy. And, if this is so, then one might argue that the Principle of Respect for Autonomy, which has so long dominated bioethical discussions, should be replaced by a focus on the Principle of Beneficence. But this argument would move too quickly, for, as was argued in this paper, persons could still be fully autonomous even with respect to those decisions that they made on the basis of self-deceptively engendered adaptive preferences. As such, then, whatever the merits of other criticisms of the

[17] Richardson (2001) has argued that autonomy is a content-laden concept, and so no purely procedural account of autonomy can be correct. To support his view Richardson compares two cases in which persons act on adaptive preferences: Jean La Fontaine's anthropomorphic fox, and Bully Stryver. Richardson holds that while the relevant preferences of both the fox and Stryver were procedurally identical (they were both adaptive, and they were both the result of self-deception) it is intuitively plausible to hold that only the fox's autonomy was compromised by his being motivated to act by the adaptive preferences that he held as a result of self-deception. Since this is so, argues Richardson, ascribing autonomy to a person with respect to her preferences and subsequent actions cannot be made on purely procedural grounds: It will also be based on an assessment of the value of the preference or act in question. As such, he concluded, autonomy is an intrinsically value-laden concept. In response to Richardson I argued that a proper understanding of the cases of both the fox and Stryver would hold that neither are autonomous with respect to their adaptive preferences and subsequent actions, precisely because both acted on adaptive preferences that arose through self-deception. (See Taylor 2009, 73-74) Since both of these persons thus failed to be autonomous with respect to their preferences and subsequent actions for the same procedural reason, I argued that this pair of cases did not support Richardson's view that autonomy was an intrinsically value-laden concept. This response to Richardson was mistaken, for I was mistaken to hold that the fox and Stryver suffered from a diminution in their autonomy as a result of their self-deceptively engendered adaptive preferences. However, since it is now clear that *both* of these persons were autonomous with respect to their self-deceptively engendered adaptive preferences (or, at least, that they could be so autonomous) the structure of my response, if not the details, is still sound, for there is still no justification for the differential ascriptions of autonomy to these persons that Richardson's argument rested on.

contemporary focus on the Principle of Respect for Autonomy, it should not be replaced out of concern for the prevalence of adaptive preferences within healthcare settings.[18]

References

Barnes, A.: Seeing Through Self-Deception. Cambridge University Press, New York (1997)
Baron, M.: What is Wrong with Self-Deception? In: Laughlin, B.P., Rorty, A.O. (eds.) Perspectives on Self-Deception, pp. 431–449. University of California Press, Berkeley (1988)
Christman, J.: Defending Historical Autonomy: A Reply to Professor Mele. Can J. Philos 23, 281–289 (1993)
Colburn, B.: Autonomy and Adaptive Preferences. Utilitas 23, 52–71 (2011)
Dickens, C.: A Tale of Two Cities. Dover Publications, Mineola (1998)
Elster, J.: Sour Grapes: Studies in the Subversion of Rationality. Cambridge University Press, Cambridge (1987)
Frankfurt, H.G.: Autonomy, Necessity, and Love. In: Frankfurt, H.G. (ed.) Necessity, Volition, and Love, pp. 129–141. Cambridge University Press, New York (1999)
Friedman, M.: Autonomy and the Split-Level Self. South J. Philos 24, 19–35 (1986)
Hamlyn, D.W.: Self-Deception. J. Med. Ethics 11, 210–211 (1985)
Mele, A.R.: Self-Deception Unmasked. Princeton University Press, Princeton (2001)
Paluch, S.: Self-Deception. Inquiry 10, 268–278 (1967)
Pears, D.: Self-Deceptive Belief Formation. Synthese 89, 393–405 (1991)
Richardson, H.S.: Autonomy's Many Normative Presuppositions. Am. Philos Q 38, 287–303 (2001)
Savers, G.M., Perera, S.: Withholding Life Prolonging Treatment and Self-Deception. J. Med. Ethics 28, 347–352 (2002)
Shute, N.: On the Beach. Vintage, New York (2010)
Sorenson, R.A.: Self-Deception and Scattered Events. Mind 94, 64–69 (1985)
Taylor, J.S.: How Much Understanding is Necessary for Autonomy (2013) (manuscript)
Taylor, J.S.: Practical Autonomy and Bioethics. Routledge, New York (2009)
Zimmerman, D.: Sour Grapes, Self-Abnegation and Character Building. Monist 86, 220–241 (2003)

[18] I thank Jukka Varelius and Juha Räikkä for their exceptionally helpful comments on an earlier draft of this paper.

Chapter 10
Adaptive Preferences and Self-Deception

Juha Räikkä

10.1 Introduction

Although it is generally known that both the so-called adaptive preference formation and self-deception reduce frustration caused by cognitive dissonance, there is surprisingly little discussion about the relation between adaptive preferences and self-deception.[1] Perhaps people have thought that the connection between them is so obvious that it does not deserve special attention.[2] Or perhaps they have thought that these two phenomena are so clearly unconnected that any discussion is unworthy.[3] Be that as it may, in this paper I will aim to provide an understanding of adaptive preference formation that is based on the idea that in many cases self-deception has a key role in leading a person to adaptive preferences. Whether the two phenomena are often closely connected depends of

Juha Räikkä
Department of Philosophy and Behavioural Sciences
University of Turku
Turku
Finland
e-mail: jraikka@utu.fi

[1] Cognitive dissonance and adaptive preferences, see, e.g., Elster (1983, 110); cognitive dissonance and self-deception, see, e.g., Scott-Kakures (2009, 73-106).

[2] The connection between adaptive preferences and self-deception is briefly mentioned for instance in Talbott (1995, 27), Richardson (2001, 292), and Zimmerman (2003). Zimmerman (2003, 222) says explicitly that "Adaptive preference-formation often involves a kind of self-deception, or at least a lack of self-knowledge" but he does not develop the idea. Cf. James Stacey Taylor's paper in this volume.

[3] The connection between adaptive preferences and self-deception is explicitly denied by Elster (1989, 175). The basis of the denial is the idea that while adaptive preference formation is a causal phenomenon, self-deception is intentional in the sense that it "has an intentional component in that I know the truth of what I am trying to hide from sight".

J. Räikkä and J. Varelius (Eds.): *Adaptation and Autonomy*, SAPERE 10, pp. 149–165.
DOI: 10.1007/978-3-642-38376-2_10 © Springer-Verlag Berlin Heidelberg 2013

course on the descriptions we give them. Therefore it should not be a surprise that a case can be made for the claim that often they are connected to each other. However, my aim here is to argue that adaptive preference formation and self-deception are often closely connected under relatively plausible descriptions. This result, if correct, seems to reveal something important about the nature of adaptive preferences. Although self-deception is in all likelihood an even more mysterious phenomenon than adaptive preference formation, the notion of self-deception may help to clarify the process of adaptive preference formation, or at least its usual instances.

I will first give an account of what I mean by adaptive preference formation. Following many others, I will rely on the general idea that a person forms her preferences adaptively as she changes her preferences merely because she notices that a desirable option is impossible or difficult to achieve. Such a process is irrational, as limited feasibility of an option does not justify downgrading its worth (or upgrading the worth of alternative options), although it may of course justify a decision not to pursue it.[4] I will then argue that usually adaptive preference formation involves a factual mistake and that it is natural to interpret the factual mistake involved as a form of self-deception, especially if we assume that self-deceivers have mistaken beliefs. Such an assumption is not completely innocent, but arguably it is not particularly problematic either. I will end my discussion by considering the normative question of how we should treat adaptive preferences, especially if they are seen to evolve from self-deception, although I will not claim that the connection to self-deception commits us to treat adaptive preferences in some new revolutionary way. Finally I will also connect my analysis to the general topic of the present volume, namely life and death choices.

10.2 Adaptive Preferences in the Light of Two Examples

The concept of adaptive preference formation is used in various ways.[5] The understanding of adaptive preferences adopted in this paper is based on the assumption that a person who forms her preferences adaptively either changes the criteria by which she evaluates the alternative options or changes her mind on the question of how well alternative options meet the criteria she uses (cf. Cohen 1995, 254-255, fn. 17). Consider two examples.

A person would like to buy an apartment in the centre of town. She prefers an apartment in the central area to an alternative apartment, located in a suburban area, close to nature trails, as the apartment in the central area is near her

[4] For some reason Elster (1983, 119) argues that "whereas adaptive preferences typically take the form of downgrading the inaccessible options, deliberate character planning would tend to upgrade the accessible ones."

[5] Recent works on adaptive preference formation include Bruckner (2009), Breakey (2010), Colburn (2011).

workplace and good restaurants. However, she notices that the apartment in the central area is far too expensive for her. Within a day or two her preference changes and she starts to prefer the suburban apartment to the apartment in the central area. She thinks that the "apartment in the central area is not worth having anyway".[6] When asked why she now prefers a suburban apartment rather than the one she just praised, she replies that it is good to have an apartment that is located close to nature trails as "Nature is extremely important to me". This claim, however, *is not true*. She has always liked good restaurants and lively city life. She has *changed the criterion* by which she compares apartments. The unobtainability of the apartment in the central area has caused her to change her preference and explains why she now believes that the suburban apartment is better than the one in the central area. If she were to win in a lottery, she would immediately change her preference to the original one and buy the apartment in the central area. Her friends know this very well but she does not.

Anne would like to date Luca who is a fair and handsome guy, i.e., in Anne's view, and he is well-known for his skills as a team leader in the big company where he works. However, Luca is not interested in Anne and he makes this clear – of course in a polite way that is so typical of his character. Suddenly Anne changes her preference and starts to prefer Michael to Luca. Michael, who is also a team leader by profession, has always liked Anne very much. When asked why she changed her mind she replies that, actually, Michael is a more interesting man than Luca, a better one in many senses. Michael is fairer, more handsome and a more skillful team leader than Luca, she says. However, the claim *is not true*. It is quite obvious that Luca *meets the criteria* Anne mentions much better than Michael, although the criteria are to a certain extent subjective. The fact that Luca is not interested in dating Anne has caused her to change her preference and explains why she believes that Michael is a better man than Luca. If Luca were to change his mind concerning Anne, she would change her mind concerning the superiority between Luca and Michael. Her friends know this very well but she does not.

These examples suggest that usual cases of adaptive preference formation have some features in common. I would like to mention the following six. First, adaptive preference formation is usually a *process* that involves a change of preference (or its strength). If the change condition is not fulfilled, we should not usually talk about adaptive preferences.[7] Second, adaptive preference formation is irrational. As already indicated, unavailability of an option does not justify

[6] Cf. Elster's (1983, 123) example.

[7] I admit that there may be exceptions. Suppose that a person prefers A to B, but then finds a reason to change her preference. At the same time, however, she notices that B is not available and *because of that* continues to prefer A to B. Now her preference *is* adaptive, although her preference never changed.

downgrading its worth (or upgrading the worth of alternative options), although it may justify a decision not to try to achieve it.[8] Third, at least in the usual instances of adaptive preference formation, the agent would change her preference to the original one if she noticed that the option she originally preferred is now available. There are exceptions, of course, because people may find good reasons for the preferences that were initially formed on irrational grounds. Fourth, an agent who forms her preferences adaptively is unaware of the fact that she is doing so and that her preferences are adaptive. This is not to say that people cannot be well aware that many of their preferences are caused by upbringing and socialization, without proper reflection (cf. Bruckner 2009, 316-317). Fifth, adaptive preferences are exceptional. In principle, it may be that *all* our preferences are caused by reasons that differ from the reasons by which we aim to justify our preferences, but even if this were so, we would not *attribute* adaptive preferences to everyone who has preferences (cf. Sie 2009, 520). Finally, adaptive preference formation involves usually a *factual mistake*.

The last condition is critical. Although both Jon Elster and Gerald Cohen, to both of whom I am greatly indebted, seem to argue that a factual mistake is involved only in examples that follow the logic of the dating example above, it seems quite clear that a factual mistake is involved also in cases that follow the logic of the apartment example (Elster 1983, 123; Cohen 1995, 254-255, fn. 17). When a person says "Nature is extremely important for me" she makes a false claim, as it is not the case that nature is extremely important for her. When people change the criteria they use, the change is usually *based on* a factual (but false) claim. In his paper on the coherence of the "self" William J. Talbott argues, echoing both Elster and Cohen, that we should actually make a distinction between adaptive preference formation and adaptive belief formation (Elster 1983, 123 cf. 109-110; Cohen 1995, 254-255, fn. 17). By "adaptive preference formation" Talbott refers to changes in preferences "that cannot be attributed to mere changes in belief – for example, the process by which most people acquire a taste for their own national or ethnic cuisine"(Talbott 1995, 27; cf. Bruckner 2009, 316). In my view, most cases of changes of preferences in adaptive preference formation can naturally be attributed to "mere changes in belief", and I doubt whether adaptive preferences would have gathered much attention were they reducible to preferences such as people's preference for their own ethnic cuisine. Unlike a preference to eat pizza or mashed potatoes more often rather than extremely hot Korean noodle dishes, adaptive preferences seem to be *reason-responsive* in the sense that in adaptive preference formation people usually feel that they have a *reason* to change their preferences – whether or not they are willing or able to explain their reasons.

[8] A person should *believe* that the option is unavailable (or that the option is very difficult to obtain).

Of course, the concept of adaptive preference formation is used in various ways, and my intention here is not to say that the understanding adopted here is superior to all others. However, the description presented above has some important features that certain other descriptions do not have. The features I have in mind are that (1) adaptive preference formation is an instance of human irrationality, and that (2) adaptive preference formation can be understood without reference to oppression or injustice or other moral concepts. In Elster's seminal contribution *Sour Grapes* (1983) he maintained that adaptive preference formation was a form of irrationality, and his world-famous example (derived from Aesop) of the fox who claims that "Those grapes are sour anyway" seems morally unproblematic.[9]

Martha Nussbaum argues that "Elster's account of adaption [of preferences] is rather narrowly focused" and, in comparison to her own account, this is definitively true (Nussbaum 2001, 78). In the view of Nussbaum (and of many others) a person has adaptive preferences, for instance, if she has preferences that she would not have had she lived in better social circumstances that would not have "caused" her to prefer things that are actually quite bad for her. An example is a woman from Bangladesh who prefers not to be educated mainly because she has lived in a society that has made her to think in the way she does.[10] As Nussbaum of course knows, this way of using the notion of adaptive preferences refers to social injustices discussed by John Stuart Mill in his 1869 classic *The Subjection of Women* in which he argues that men enslave women's minds.[11] I have no objections to the practice of using the concept of "adaptive preference formation" in this way – regardless of the fact that Mill did not use that term – but it is good to have two things in mind if one adopts this practice.[12]

First, if a person who thinks that she does not need education has her own reasons for this conviction, she need not be in any sense *irrational*, even if some other people could plausibly claim that her conviction is not very clever and that it is due to oppressive social circumstances, or even that we need not take her conviction into account in political decision-making. Indeed, she may be an exceptionally rational person. Second, this way of using the notion of adaptive preferences seems to rely on the assumption that we can always distinguish circumstances that have an oppressive or unjust nature from those that have not – otherwise we would probably end up in a position that all preferences are always "adaptive", as it is likely that nobody is completely free of the external effects on one's preferences (cf. Walker 1995, 458-459). Perhaps such a distinction can be made, or at least we may have some clear cases of oppressive and unjust

[9] In the fable of The Fox and the Grapes, the fox seems to have a mistaken belief because of a *mistaken perception*, caused by a desire. For a discussion, see e.g. Balcetis and Dunning (2006, 2010).

[10] As explained by Walker (1995, 461) who refers to Nussbaum.

[11] Nussbaum (2001, 80) refers to Mill.

[12] For an interesting discussion on the positions taken by Mill and Nussbaum, see, e.g., Khader (2011, 31 and 47). See also Baber (2007).

circumstances. As Nussbaum herself points out in her *Creating Capabilities* (2011), "It is not possible to identify the preferences that represent adaptation […] without an independent theory of social justice" (Nussbaum 2011, 84). It is unclear whether all who have adopted the wide understanding of adaptive preference formations have such an independent theory.[13]

10.3 What Happens in Self-Deception?

I will now turn to the notion of self-deception. The question of what happens in self-deception is quite difficult, and I assume that nobody knows the answer. However, here I will briefly introduce Alfred Mele's theory as presented in *Self-Deception Unmasked* (2001) and developed for instance in his recent paper on "When Are We Self-Deceived?" (2012). The point here is not to claim that Mele's *explanation* of self-deception is the final truth of the matter (or even close to it), but only to show that self-deception can be quite plausibly *described* in a way which suggests that it need not be as mysterious and paradoxical a phenomenon as many seem to have thought. In *Sour Grapes* Elster explicitly argues that, because of the paradoxical nature of self-deception, the cases that "superficially" look "like self-deception" (Elster 1983, 149) are in fact cases of "wishful thinking" (Elster 1983, 152).[14] Many would agree that Elster exaggerated the complexity of the notion of self-deception and that there is no need to be skeptical about the existence of the phenomenon.

According to Mele, beliefs based on self-deception are motivationally biased. In ordinary ("garden-variety") cases of self-deception people believe something that they would like to be true. Mele mentions beliefs such as that they are not seriously ill, that their children are not addicted to drugs, and that their spouses do not have affairs with anyone (Mele 2001, 4, 57). Self-deception is an explanatory concept:

> We postulate self-deception in particular cases to explain data: for example, the fact that there are excellent grounds for holding that S believes that *p* despite its being the case that evidence S possesses makes it quite likely that ~*p*. And we should ask how self-deception is likely to be constituted – what it is likely to be – if it does help to explain the relevant data. (Mele 2001, 10)

In Mele's view, self-deception does not involve an agent's intention or an attempt to deceive oneself (although other kinds of intentions are often involved) and he

[13] No doubt, those who have adopted the wide understanding of adaptive preference formations need not have an independent theory of justice, if they concentrate merely on clear cases.

[14] Mele (2001, 73-74) criticizes explicitly the idea that self-deception could be reduced to wishful thinking.

defends the doctrine he calls the "antiagency view" (2001, e.g. 13, 18, 49).[15] The challenge of the antiagency view is to provide a plausible account of the mechanisms "by which desires lead to motivationally biased beliefs", and providing such an account is Mele's main project (Mele 2001, 14). He does not deny that there could be "hidden intentions" or something similar, but he argues that such notions do not provide the *best explanation* for the ordinary cases of self-deception (Mele 2001, 17, 23-24). Self-deception should not be understood on the basis of interpersonal model of deception that (usually) involves intention to deceive (Mele 2001, 25). Rather, the question is about unintentional but motivated *data manipulation*. The idea that self-deception involves mistreatment of data is of course a part of virtually every theory of self-deception. However, Mele offers four rather detailed examples (that are not meant to be exhaustive) of how our desiring that *p* can contribute to our believing that *p*. First, in *negative misinterpretation* our "desiring that *p* may lead us to misinterpret as not counting (or not counting strongly) against *p* data that we would easily recognize to count (or count strongly) against *p* in the desire's absence". Second, in *positive misinterpretation* our "desiring that *p* may lead us to interpret as *supporting p* data that we would easily recognize to count against *p* in the desire's absence". Third, in *selective focusing/attending* our "desiring that *p* may lead us both to fail to focus attention on evidence that counts against *p* and to focus instead on evidence suggestive of *p*". Fourth, in *selective evidence-gathering* our "desiring that *p* may lead us both to overlook easily obtainable evidence for *~p* and to find evidence for *p* that is much less accessible" (Mele 2001, 26-27). Mele stresses that in none of the cases does an agent intend to deceive oneself, although these activities may very well be involved in ordinary cases of self-deception. People's beliefs can be biased because of many reasons (such as vividness of evidence) but in ordinary cases of self-deception the biasing is *motivated* (Mele 2001, 28).

In "When Are We Self-Deceived?" Mele argues that there are six conditions that are "jointly sufficient for entering self-deception in acquiring a belief that p" (Mele 2012, 12). The conditions he has in mind are the following. First, the "belief that *p* which S acquires is false". Second, "S treats data relevant, or at least seemingly relevant, to the truth value of *p* in a motivationally biased way". Third, this "biased treatment is a nondeviant cause of S's acquiring the belief that *p*".[16] Fourth, the "body of data possessed by S at the time provides greater warrant for *~p* than for *p*".[17] Fifth, "S consciously believes at the time that there is a significant chance that ~p". Sixth, "S's acquiring the belief that p is a product of 'reflective, critical reasoning,' and S is wrong in regarding that reasoning as

[15] Mele (2001, e.g. 16) points out that it is unproblematic that a person can intentionally deceive himself, but that these cases do not represent garden-variety cases of self-deception.

[16] Mele (2001, 51) points out that the term "nondeviant" in the third condition is problematic and that specifying "the precise nature of nondeviant causation of a belief by motivationally biased treatment of data is a difficult technical task".

[17] Mele (2001, 50-51) points out that the fourth condition is not necessary.

properly directed" (Mele 2012, 12). Mele is well aware that many philosophers and psychologists have emphasized that a sort of *tension* felt by a person is a crucial feature of self-deception, but he points out that the conditions do not rule out that a self-deceiver feels tension. It is just that there are cases of (garden-variety) self-deception in which such tension is absent (Mele 2001, 52).

Notice that not all ordinary cases of self-deception are those in which a person *acquires* a new belief. There are cases of ordinary self-deception in which a person *retains* an old belief but still drifts into self-deception. Mele's example is Sam who has believed for years that his wife, Sally, would never have an affair with another man (Mele 2001, 56-57). This belief has been warranted for a long period of time but, as it happens, it is no longer warranted, as Sally *is having* an affair. However, despite the evidence Sam possesses – "Sally is now arriving home late from work on the average of two nights a week, she frequently finds excuses to leave the house alone after dinner and on weekends", and so on – he refuses to change his belief and drifts into self-deception. According to Mele, the process here is similar to that of entering self-deception in acquiring a new belief. "Each of the types of data manipulation [...] may occur in a case of this kind" (Mele 2001, 57).[18]

Entering self-deception should be distinguished from remaining self-deceived. Mele's theory concentrates on the former issue, but it can possibly be applied to the latter as well. According to Mele (and many others) a person who enters self-deception does not first hold a certain view (say that $\sim p$) and then change it (to p). Rather, in self-deception the person has not made up her mind regarding the issue whether p, and then (by unintentionally manipulating data) ends up with the view which she wants to be true (except in cases of "twisted" self-deception).[19] Possibly, at least in some cases, remaining in self-deception involves "repeated resistance to the truth", as argued by Kent Bach, which would mean that the question whether p keeps appearing constantly in the person's consciousness (Bach 2002, 204). Remaining in self-deception would therefore imply that a person must constantly enter self-deception.[20] To an extent, this sounds plausible, and might explain feelings of tension that are quite commonly seen as an essential part of being self-deceived.

An interesting question is why some people with certain desires enter into self-deception while others with similar desires do not. An equally interesting question is why a person who lets her desires influence her beliefs on certain issues is able to resist the temptation in other occasions. More generally, it would be interesting

---

[18] It is tempting to think that *at some level* Sam must believe that the data he manipulates is aversive (this is why he manipulates it in the first place) and that it must be aversive just because it is evidence for adultery. If so, then at some level Sam seems to know the truth. See, e.g., Lockie (2003, 299).

[19] In "twisted" self-deception a person ends up with a belief which she wants to be *false*, even if the evidence points to a different direction (Mele 2001, Ch. 5).

[20] Bach (2002, 204) emphasizes that self-deception should be distinguished from "tension-free dogmatism."

to know how to explain the apparently strategic behavior (i.e. data manipulation) of self-deceivers, given that they do not intend to deceive themselves. Very roughly, Mele's reply (that is based on earlier work by James Friedrich, Yaacov Trope and Akiva Liberman) is that a person who ponders whether it is the case that *p* and tests her hypothesis (that *p*) tries to avoid serious *error costs*, i.e. expected costs caused by believing that p when it is not the case that *p* and expected costs caused by believing that ~*p* when it is the case that *p* (Mele 2001, 31-46). A crucial point here is that error costs can be *asymmetric* in the sense that, in a given situation, it may cause more costs for the person to believe falsely that *p* than to believe falsely that ~*p*. When error costs are asymmetric, the person can adopt a different *confidence threshold* for accepting her hypothesis and for rejecting her hypothesis.[21] The basic principle is that the "lower the threshold, the thinner the evidence required to reach it" (Mele 2001, 34). Obviously, confidence thresholds also determine how long the person is willing to test her hypothesis before drawing her conclusions. The outcome of all this is data manipulation and biased hypothesis testing which leads the person to mistaken beliefs that are consistent with her desires, i.e. to self-deception (cf. Scott-Kakures 2002, 698).[22]

However, for the present purposes, it is irrelevant whether the antiagency view holds and whether Mele's six conditions introduced above are the best possible ones. However, I do assume that (1) self-deception does involve data manipulation of the types that Mele mentions and that (2) the first of Mele's six conditions that are in his view jointly sufficient for entering self-deception in acquiring a belief that *p*, namely the False Belief Condition, is correct in the sense that deceiving oneself entails having a false belief.[23] The False Belief Condition (which says that the belief that *p* which self-deceiver S acquires is *false*) is crucial, as my intention

[21] For a discussion about confidence thresholds, see, e.g., Ellees (1983), Trope, Gervey, and Liberman (1997), Hawthorne and Bovens (1999).

[22] Mele (2001, 38, 48) emphasizes that that he is not concerned in defending the pragmatic model of hypothesis testing "in its full generality." This is a clever move from him, to say the least. It seems clear that usually our motivation in testing hypotheses is related to desire to *know* whether it is the case that *p* (which is an epistemic motivation) rather than motivation to avoiding costly errors. As pointed out by Scott-Kakures, the pragmatic model that Mele applies does not say a word about the questions such as why one motivation is currently at work rather than the alternative one and why people do not realize on which motivation they are relying now when they have taken the crucial step (i.e. the choice between alternative motivations). Dion Scott-Kakures has presented an interesting alternative, based on Leon Festinger's well-known theory of cognitive dissonance, and in Scott-Kakures' view what "motivates inquiry is, through and through, an epistemic motive" (Scott-Kakures 2009, 105). If so, then the question "Why is one motivation currently at work rather than the alternative one" is pointless.

[23] The concept of False Belief Condition derives from Kevin Lynch (2010). Lynch's paper is a reaction to my previous work on the topic of self-deception (see Räikkä 2007). Of course, Mele's list of *sufficient* conditions does not commit him to accept False Belief Condition, but he does think that self-deceivers have false beliefs, see Mele (2010, 745-750, 747 fn. 3, 748 fn. 4).

here is to argue that it is natural to interpret the factual mistake which is often involved in adaptive preference formation as a form of self-deception. If self-deception does not involve false beliefs (i.e. factual mistakes) then such connection may be difficult to establish. It may seem that the False Belief Condition is unproblematic, as it is involved in many theories of self-deception even when those theories are otherwise very different both from each other and from Mele's theory. It is not hard to guess that the False Belief Condition is popular mainly because of purely lexical reasons. As Mele writes, "A person is, by definition, *deceived in* believing that *p* only if *p* is *false*; the same is true of being *self-deceived* in believing that *p*" (Mele 2001, 51).[24] However, it is important to notice two points here.

First, a person can be deceived *into* believing something which is *true*. Mele (2001, 129) makes this point very clear in *Self-Deception Unmasked*, but the following example comes from Kevin Lynch (2010, 1074):

> Say that I know that *p* is true and I know you will only believe that *p* if you hear it said from Jones, who is the only man you trust on this matter. I then lie to you by saying that I was in touch with Jones, who told me that *p*, and you then believe that *p*. In this situation, you are not *deceived in* believing that *p*, since this is a true belief. However, we can say here that you were *deceived into* believing this true proposition by me.

Perhaps a person can be similarly self-deceived *into* believing something which is true (although I have my doubts). In any case, the discussion here concerns cases in which a person is self-deceived *in* believing something.

Second, there are theories of self-deception that do not involve or that explicitly reject the False Belief Condition. For instance, we could distinguish between what a person *actually* believes and what she *takes herself to believe* and argue that in self-deception people's actual beliefs are *true*.[25] Here is an example.[26] Amy takes herself to believe that it is not the case that her daughter Bea is anorexic (although Bea evidently is anorexic and Amy possesses all the evidence needed) but *actually* Amy believes that Bea *is* anorexic.[27] That explains why Amy's talk and overall

---

[24] Suppose that a person treats data in a motivationally biased way and concludes that *p* – despite its being the case that the evidence the person possesses makes it plausible that it is not the case that *p*. Suppose also, however, that it happens to be the case that *p*. The question is not about literal self-*deception*, although we may say that something went wrong.

[25] For a distinction between what a person believes and what she takes herself to believe, see e.g. Moran (2005). For a discussion, see, e.g., Räikkä and Smilansky (2012).

[26] Cf. Bagnoli's (2012, 99-100) example.

[27] Georges Ray (1988) has distinguished between what he calls "central beliefs" and "avowals". See also Audi (2007). As is well known, Audi has defended the idea that self-deceivers' do not have false beliefs.

behavior seem to be somewhat incoherent.[28] I admit that my argument here is not consistent with theories such as this, although it does not follow that a close connection between adaptive preference formation and self-deception could not still be established. It only follows that the connection should be established in some way other than by using the False Belief Condition as a vehicle.

10.4 Self-Deception as a Basis of Adaptive Preference Formation

Let us consider again the apartment example and the dating example that represent typical instances of adaptive preference formation (as understood here). Can these cases be explained by referring to self-deception? In this section I will argue that they can. The idea that self-deception involves mistreatment of data is a part of any sensible theory of self-deception, and both the apartment example and the dating example can plausibly be described as cases of data manipulation.

Consider the dating example first. Anne notices that Luca is not interested in her and she starts to believe, *falsely*, that Michael is fairer, more handsome and a more skillful team leader than Luca. This false belief helps her to adopt a new preference which is quite obviously an adaptive preference.[29] But how *can* she acquire this new belief which looks unwarranted? Plausibly, the question is about self-deception. In the circumstances described, Anne *wants* it to be true that Michael is fairer, more handsome and more skillful than Luca, and this leads her to data manipulation. Unlike Luca, Michael has a bit of a bad reputation, but in Anne's view this is merely because people around tend to envy Michael (a negative misinterpretation of data). Unlike Luca, Michael quarrels with his team members every now and then, but in Anne's view this is a sign that Michael has courage that team leaders need (positive misinterpretation of data).

Let us now turn to the apartment example. A person noticing that the apartment in the centre of town is far too expensive for her starts to believe, *falsely*, that nature is very important to her. The question is how *can* the person acquire this new belief which must look unwarranted? In all likelihood, the person is deceiving herself. In the unfortunate circumstances the person wants it to be true that nature is very important to her, and this makes her to manipulate data. She knows that she has spent a lot of time in restaurants and coffee houses in the centre of town, but she also knows that every now and then she has visited the countryside, and she draws conclusions from this latter fact rather than the former (*selective focusing/attending*). When she thinks what kind of attitude she has to nature (and

[28] The idea that self-deceiver's talks and overall behavior can be incoherent is emphasized in Rorty (1988).

[29] Following Ben Colburn (2011, 66), we may want to speak of "covert influences", but it does not seem to be the case that a person's reason for her new preference is covert (if something is). The reason (i.e. a belief) is merely false, and obviously she does not know that it is false. For a criticism, see Bruckner (2011).

nature trails in particular) she finds more and more evidence for the view that actually she loves to spend time there, seeing birds and old trees, as well as meeting joggers (*selective evidence-gathering*).

Why does Anne not deceive herself by believing, falsely, that actually Luca *is* interested in her? That would be a rather unsurprising way to drift into self-deception, i.e. if she is to drift into self-deception in the first place. And why does not the person who aims to buy an apartment deceive herself by believing, falsely, that the apartment in the central area is *not* too expensive for her? Again, that would be an unsurprising way to deceive oneself. A plausible reason why the persons in both examples deceive themselves in the way they do is that the alternative ways may be psychologically impossible for them. Perhaps the evidence shows too clearly that Luca is not interested in Anne, so that Anne simply *cannot* think that he does. Similarly, perhaps the evidence shows too obviously that the apartment in the central area is too expensive, so that the person who aims to buy an apartment there just *cannot* think otherwise. As many authors have pointed out, although self-deceivers seem to be able to believe almost everything, there are upper limits on what kinds of lies to oneself are possible – in usual (garden-variety) self-deception (Mele 2001, 57).

In my view, the dating example is clearly a case of adaptive preference formation, and under the given description it seems to be also a case of self-deception. Similarly, I think that the apartment example is a case of adaptive preference formation (as argued in section 10.2), and, plausibly, it is also a case of self-deception. Someone might object that in usual self-deception beliefs that are accepted contrary to the overall evidence are beliefs that cause pleasure (or reduce frustration) almost *immediately* – beliefs such as "I am not seriously ill" or "My child is not addicted to drugs". The belief that "Michael is more handsome than Luca" does not seem to be a belief of that type, neither does the belief that "Nature is very important for me". But this objection is flawed. It is likely that the main function of self-deception is to reduce cognitive dissonance and psychological discomfort associated with it, and this function can easily be fulfilled even if the accepted (false) beliefs need further thoughts before the relief comes. For instance, in the apartment example the (false) belief that "Nature is very important for me" *does* reduce frustration, as the person who believes so has appropriate other beliefs.

10.5 Some Normative Implications

By way of presenting examples I have argued that adaptive preference formation and self-deception can be closely connected. I do not suggest that two examples show that they are often connected in the way I introduced them, but it would not be difficult to come up with similar examples. It seems to me that adaptive preference formation and self-deception usually have quite a close connection. In this section I will ask how we should treat adaptive preferences that evolve from self-deception. The ethics of self-deception is a separate question that falls beyond

the scope of this paper.[30] Here my concern is simply whether we should somehow *disqualify* adaptive preferences and in particular the decisions based on such preferences.

To start, it seems clear that in general we should deal with people's beliefs that are actually based on self-deception as if those beliefs were their normal beliefs, even if we may not be happy that the beliefs are based on self-deception and even if in some circumstances we are permitted or even obligated to try to *free* people from self-deception (as this may cause serious harm to other people). Similarly, in usual cases we should treat people's preferences that are actually adaptive as if they were their normal preferences. Suppose that Anne and Michael get married. It is quite clear that we should treat their marriage both legally and morally just like any other marriage – whether or not we think that something went wrong because Anne's picture of Michael was based on self-deception. We may want to say that, because of Anne's self-deception, she was not ideally autonomous or that there was a "failure of authorship", as Carla Bagnoli puts it (Bagnoli 2012, 111). But even if this were so it would be out of the question to claim that their marriage was not legally binding or that the marriage was *exceptionally* problematic in moral terms.[31] Other things equal, people are free to make stupid decisions, as Anne possibly did. Even if Anne's preference ("Michael is better than Luca") was adaptive, we ought to treat that preference as though it were her normal preference, even if we may not be happy that the preference was clearly adaptive.

Suppose that someone buys an apartment in the suburban area, close to nature trails, and it turns out that the decision was partly due to self-deception. The buyer falsely believed that nature is very important to her. When she finally finds out the truth – that nature is not very important to her – could she inform the vendor that in fact she had deceived herself and therefore the contract was illegal? Of course not. Whether or not her decision was partly due to self-deception, the contract was legally binding, as the self-deception does not diminish autonomy in the sense that it would be relevant in the context. Even if the buyer's preference ("The suburban apartment is better than the apartment in the centre of town") was adaptive, it should be treated as though it were her usual preference, i.e. it should *not* be disqualified.

It may be tempting to think that adaptive preferences should *always* be treated as if they were usual preferences. But obviously this is not the case. The topic of the present volume is, *inter alia*, "life and death choices", and in this context it is far from clear how adaptive preferences should be treated. As opposed to marriages and decisions concerning where to live, life and death choices are or at least can be, by definition, *final choices*. Consider the Diane Pretty case, as described by S.B. Chetwynd (2004, 173):

---

[30] The ethics of self-deception, see, e.g., Martin (1986), Baron (1988), Barnes (1997), Levy (2004).

[31] No doubt, *before* the marriage, a close friend may want to warn Anne that she is deceiving herself.

> Diane Pretty was a 42-year-old woman in the advanced stages of motor-neurone disease. The symptoms of the disease include gradual paralysis leading to eventual loss of respiratory muscle function and respiratory failure, and, because Mrs. Pretty suffered this disease in its bulbar from, included swallowing difficulties. Wanting to pre-empt a possibly painful and distressing death through choking and asphyxia, she secured her husband's agreement to assist her suicide, provided he wouldn't fall foul of the Suicide Act.[32]

Suppose (probably contrary to the facts) that Diane Pretty's preference to die was adaptive, and that the adaptive preference was caused by self-deception. Should we say that her preference should be seen as if it had been a normal preference? That is, should we say that everything went more or less correctly even if there was a minor problem involved, namely self-deception that helped her to form adaptive preference? I do not think so. The consequences of the final choices are or at least can be so dramatic that "neutral orientation" which might be appropriate in many or even most contexts seems to be out of place. Here disqualification is a relevant option. Notice that this does not depend on whether we consider adaptive preferences *before* the decisions or *after* the decisions. In the context of the final choices adaptive preferences seem quite problematic in both cases. How much rationality (or autonomy) we require from the decision-makers depends upon the context, and life and death choices are such that acceptable rationality requirement is quite demanding. The choice of death over life is something very different than a choice of Coke over Pepsi, as in the latter case we probably do not demand the decision-maker to be particularly rational (cf. Walker 1995, 471, fn. 14). Whether self-deception is bad for autonomy is a separate question, but surely an argument can be made that it is bad. Self-deceivers do not seem to have "relevant beliefs about the nature and consequences of their actions", and if autonomy requires such beliefs (as argued by Tom Beauchamp and James Childress (2009, 127)), then, possibly, self-deceivers are not autonomous.

What is the upshot of all this? The main message is rather moderate. Although in usual cases we ought to deal with people's beliefs that are actually based on self-deception as if those beliefs were their usual beliefs, and also to treat people's preferences that are actually adaptive as if they were their usual preferences, this need not hold when we are concerned about the final choices. The lack of an analogy between "usual cases" (such as marriages) and the final choices is simply too strong. On the other hand, as far as self-deception and adaptive preferences are linked in the way suggested above, then there does not seem to be a reason to think that a similar link could not be found, at least in some cases, when the decisions concern life and death. Therefore, the "usual cases" may very well provide us information about the process of adaptive preference formation in the life and death choices.

[32] For similar examples, see, e.g., Grey (1999), Guerra (1999).

10.6 Concluding Remarks

I would like to end my discussion by making two points that may sound obvious but are in my view so important that they are worth stating.

First, patients with terminal illness get most of the relevant information concerning their condition from medical professionals. Their words and reports strongly influence the patient's view about issues such as what are the chances, how painful the rest of the life will be, how long it might be, and so on. Since we know that diminishing options increases the possibility of adaptive preference formation, medical professionals should be extremely careful in order not to give an incentive for such preference formation. But it is anything but easy to avoid it, if the professionals do not want to mislead the patient, as they of course should not. As far as adaptive preference formation evolves from self-deception, the patients should be warned against deceiving themselves. But the whole idea of providing such warnings sounds suspicious, at least if it is not done with special sensitivity.

Second, identifying adaptive preferences is a complicated matter, and suspecting patients with terminal illness of having irrational preferences can be ethically problematic. We may think that rapid changes of preferences after bad news are a sign that indicates that preferences are adaptive. But obviously a person can change her preferences very quickly without any irrationality. Preferences are like beliefs in the sense that we cannot simply decide what to believe or what to prefer.[33] We need reasons. The idea that bad news cannot provide appropriate reasons to change one's preference is mistaken, although unavailability of an option – say, life – does not *in itself* justify downgrading its worth. In theory, medical professionals should *ask* what the reason behind the new preference is, and estimate its plausibility given the patient's other beliefs and values, keeping in mind the possibility of self-deception. But again, the whole idea of such cross-examination sounds bad.

Acknowledgements. I would like to thank Carla Bagnoli, Alfred Mele, Olli Koistinen and Jukka Varelius for helpful comments.

References

Audi, R.: Belief, Intention, and Reasons for Action. In: Timmons, M., Creco, J., Mele, A. (eds.) Rationality and the Good: Critical Essays on the Ethics and Epistemology of Robert Audi, pp. 248–259. Oxford University Press, New York (2007)

Baber, H.E.: Adaptive Preferences. Soc. Theory Pract. 33, 105–126 (2007)

Bach, K.: Book Review of Self-Deception Unmasked. Philos Psychol. 15, 203–206 (2002)

Bagnoli, C.: Self-Deception and Agential Authority: A Constituvist Account. Humanamente, 99–116 (2012)

[33] Barry (2001, 36).

Balcetis, E., Dunning, D.: Wishful Seeing: Desired Objects are Seen as Closer. Psychol. Sci. 21, 147–152 (2010)
Balcetis, E., Dunning, D.: See What You Want to See: Motivational Influences on Visual Perception. J. Pers. Soc. Psychol. 91, 612–625 (2006)
Barnes, A.: Seeing through Self-Deception. Cambridge University Press, Cambridge (1997)
Baron, M.: What is Wrong with Self-Deception? In: Laughlin, B.P., Rorty, A.O. (eds.) Perspectives on Self-Deception, pp. 431–449. University of California Press, Berkeley (1988)
Barry, B.: Culture and Equality. Polity Press, Cambridge (2001)
Beauchamp, T.L., Childress, J.F.: Principles of Biomedical Ethics, 6th edn. Oxford University Press, New York (2009)
Breakey, H.: Adaptive Preferences and the Hellenistic Insight. Aust. J. Prof. Appl. Ethics 12, 29–39 (2010)
Bruckner, D.W.: Colburn on Covert Influences. Utilitas 23, 451–457 (2011)
Bruckner, D.W.: In Defense of Adaptive Preferences. Philos Stud. 142, 307–324 (2009)
Chetwynd, S.B.: Right to Life, Right to Die and Assisted Suicide. J. Appl. Philos 21, 173–182 (2004)
Cohen, G.A.: Self-Ownership, Freedom, and Equality. Cambridge University Press, Cambridge (1995)
Colburn, B.: Autonomy and Adaptive Preferences. Utilitas 23, 52–71 (2011)
Ellees, E.: On a Recent Theory of Rational Acceptance. Philos Stud. 44, 331–343 (1983)
Elster, J.: Sour Grapes – Utilitarianism and the Genesis of Wants. In: Christman, J. (ed.) The Inner Citadel, pp. 170–188. Oxford University Press, New York (1989)
Elster, J.: Sour Grapes: Studies in the Subversion of Rationality. Cambridge University Press, Cambridge (1983)
Grey, W.: Right to Die or Duty to Live? The Problem of Euthanasia. J. Appl. Philos 16, 19–32 (1999)
Guerra, M.J.: Euthanasia in Spain: The Public Debate After Ramon Sampedro's Case. Bioethics 13, 426–432 (1999)
Hawthorne, J., Bovens, L.: The Preface, The Lottery, and The Logic of Belief. Mind 108, 241–264 (1999)
Khader, S.J.: Adaptive Preferences and Women's Empowerment. Oxford University Press, Oxford (2011)
Levy, N.: Self-Deception and Moral Responsibility. Rati 17, 294–311 (2004)
Lockie, R.: Book Review of Self-Deception Unmasked. Philos 78, 296–300 (2003)
Lynch, K.: Self-Deception, Religious Belief, and the False Belief Condition. Heythrop J. 51, 1073–1074 (2010)
Martin, M.M.: Self-Deception and Morality. University of Kansas Press, Lawrence (1986)
Mele, A.R.: When Are We Self-Deceived? Humanamente 20, 1–15 (2012)
Mele, A.R.: Approaching Self-Deception: How Robert Audi and I Part Company. Conscious Cogn. 19, 745–750 (2010)
Mele, A.R.: Self-Deception Unmasked. Princeton University Press, Princeton (2001)
Moran, R.: Problems of Sincerity. P. Aristotelian Soc. 105, 341–361 (2005)
Nussbaum, M.: Creating Capabilities: The Human Development Approach. Belknap Press, Cambridge (2011)
Nussbaum, M.: Adaptive Preferences and Women's Options. Econ. Philos 17, 67–88 (2001)
Räikkä, J., Smilansky, S.: The Ethics of Alien Attitudes. Monist 95, 511–532 (2012)
Räikkä, J.: Self-Deception and Religious Beliefs. Heythrop J. 48, 513–526 (2007)

Ray, G.: Toward a Computational Account of Akrasia and Self-Deception. In: Laughlin, B.P., Rorty, A.O. (eds.) Perspectives on Self-Deception, pp. 264–296. University of California Press, Berkeley (1988)

Richardson, H.S.: Autonomy's Many Normative Presuppositions. Am. Philos Q 38, 287–303 (2001)

Rorty, A.O.: The Deceptive Self: Liars, Layers, and Lairs. In: Laughlin, B.P., Rorty, A.O. (eds.) Perspectives on Self-Deception, pp. 11–28. University of California Press, Berkeley (1988)

Scott-Kakures, D.: Unsettling Questions: Cognitive Dissonance in Self-Deception. Soc. Theory Pract. 35, 73–106 (2009)

Scott-Kakures, D.: Book Review of Self-Deception Unmasked. Mind 111, 696–701 (2002)

Sie, M.: Moral Agency, Conscious Control, and Deliberative Awareness. Inquiry 52, 516–531 (2009)

Talbott, W.J.: Intentional Self-Deception in a Single Coherent Self. Philos Phenomen Res. 55, 27–74 (1995)

Trope, Y., Gervey, B., Liberman, N.: Wishful Thinking from a Pragmatic Hypothesis-Testing Perspective. In: Myslobodsky, M. (ed.) The Mythomanias: The Nature of Deception and Self-Deception, pp. 105–131. Lawrence Erlbaum, Mahwah (1997)

Walker, J.D.: Liberalism, Consent, and the Problem of Adaptive Preferences. Soc. Theory Pract. 21, 457–471 (1995)

Zimmerman, D.: Sour Grapes, Self-Abnegation and Charatcter Building: Non-Responsibility and Responsibility for Self-Induced Preferences. Monist 86, 220–241 (2003)

Chapter 11
Procedural Exclusion Criteria for Adaptive Preferences

Till Grüne-Yanoff and Sven Ove Hansson

11.1 Introduction

A preference is said to be adaptive if it is formed or changed in response to the agent's feasible options (Elster 1985, Bruckner 2009). Adaptive preferences are major candidates for being discounted (or entirely dismissed) in welfarist accounts of the good (Elster 1985, Bovens 1992, Nussbaum 2001). In this paper, we refine this basic intuition in two ways. First, we show that not all adaptive preferences should be thus discounted. Second, we provide a general framework for determining which preferences should be discounted due to their adaptiveness, including a set of procedural exclusion criteria to be used for this purpose.

The paper proceeds as follows. Section 2 distinguishes different kinds of exclusion criteria. Section 3 presents a number of exemplary cases, which illustrate the divergent intuitions about adaptive preferences – some preferences, because they are adaptive, intuitively should be excluded, while others, although they are adaptive, should intuitively not be excluded. Section 4 presents a formal framework with which to analyze these and other cases. Section 5 employs this framework to devise a procedural exclusion criterion, which we name *reconstructability*. Section 6 applies this formal tool to the cases of section 3. As we show, all cases can be rationally reconstructed, but only under specific kinds of frames. Section 7 presents exclusion criteria for these frames. As we show, the preferences that should be intuitively excluded are in fact excluded by these criteria. Section 8 concludes.

11.2 A Hierarchy of Exclusion Criteria

The situation we are concerned with involves an agent (the "reviewed" or the "reviewee") who holds preferences over a set of options, and another agent (the "reviewer") who seeks to determine what is best for the reviewed, taking into account

Till Grüne-Yanoff · Sven Ove Hansson
Dept. of Philosophy, Royal Institute of Technology, Stockholm, Sweden
e-mail: till.grune@abe.kth.se, soh@kth.se

J. Räikkä and J. Varelius (Eds.): *Adaptation and Autonomy*, SAPERE 10, pp. 167–182.
DOI: 10.1007/978-3-642-38376-2_11 © Springer-Verlag Berlin Heidelberg 2013

some of the reviewee's preferences, but possibly excluding others. The reviewer is typically a person or an institution who takes decisions concerning the reviewee's interest (e.g. a medical doctor, a regulator, a legal guardian), while the reviewed is typically an individual person (but might also be an institution like a company or a club, to the extent that preferences can be attributed to it).

We disagree with the claim that a preference's adaptiveness is always a good reason for it being discounted. According to our intuition, while some preferences should be excluded because of their adaptiveness (for example, Stockholm syndrome preference), other preferences might be good candidates for being accepted in a welfarist account exactly because they are adapted to the agent's circumstances. Examples of the latter include preferences adapted to one's talents and abilities.

We distinguish three kinds of arguments for discounting someone's preferences. The first of these refers to *the nature of the reviewee's preferences*. Preferences may for instance be discounted because they are inconsistent, because they do not cohere with the same person's second-order preferences, because they are unstable, etc.

The weakest criterion we refer to in this paper consists in checking for consistency of the subjective attitudes that an agent has at one point in time. It might be the case, for example, that an agent desires a particular event to happen, but wishes to avoid the consequences she knows that such an event produces. Such violations of *synchronic consistency* are perhaps the weakest exclusion criterion for adaptive preferences. A slightly stronger criterion checks the consistency of an agent's subjective attitudes across some change of background beliefs. The adaptations considered most typical – like sour grape cases – involve some learning experience that triggers a preference change. The fox, to stay with the sour grapes example, learns that he cannot reach the grapes, and consequently changes his preferences for them. *Diachronic consistency* checks whether the learning experience can rationally justify the preference change. If not, then such changed preferences should be excluded from any welfarist account of the good.

The second category are arguments that refer to the *the origin and formation of the reviewee's preferences*. Preferences can for instance be rejected because they were formed under external pressure or force, or under conditions with insufficient access to information about alternatives and background facts.

One criterion belonging to this second category exposes agents to new experiences, thus giving them the opportunity to develop new attitudes and revise their old ones, and then checks for consistency of the posterior attitudes. The Amish practice of *rumspringa* – a period during which Amish youth are permitted to depart from the strict regulations of their community – is a real-life example of such a subjective stress test of preference (Mazie 2005).

Rumspringa and similar practices might aim at reflecting on existing preferences. Alternatively, they might also aim at reflecting on the *framing of preference justifications*, in particular the number and nature of options preferences are defined over, and the background motives to which these options are tied. Indirectly, reflecting on the framing might lead to preference revision. Imposing such extra-subjective criteria on framing of preference justifications thus constitutes a more demanding review than merely exposing the reviewed to new experiences. For example, options

not under the reviewee's control should not be included among her alternative actions, while those under her control should be included; and the individuation of background conditions should be free of coercion. Imposing such a criterion clearly is stronger than merely requiring consistency. Yet it still does not refer to other persons' values concerning the same subject matter, and therefore must be distinguished from the following two, stronger, kinds of reviews.

Intersubjective tests compare attitudes of many agents in a population with different experiences. For example, a population might contain deaf (D) and hearing (H) persons, and also persons who were once hearing but now are deaf (H-D), as well as those who were deaf but now are hearing (D-H). Comparison of preferences dominant in each of these groups might give reasons for exclusion. For example, if we found that both H-D and D-H expressed a preference for hearing to being deaf, then it would perhaps seem difficult to accept a D-member's preferences for deaf to hearing.

The third category are arguments that refer to the reviewer's own values concerning the same subject matter, or to values shared or professed to be shared by a group.[1] We may for instance reject someone's preferences because we consider them to be perverse, immoral, or imprudent. Because their source is unrelated to what the reviewee values in the subject matter, we call these criteria *extra-subjective*.

The five different test criteria are depicted in figure 1, ordered into three categories.

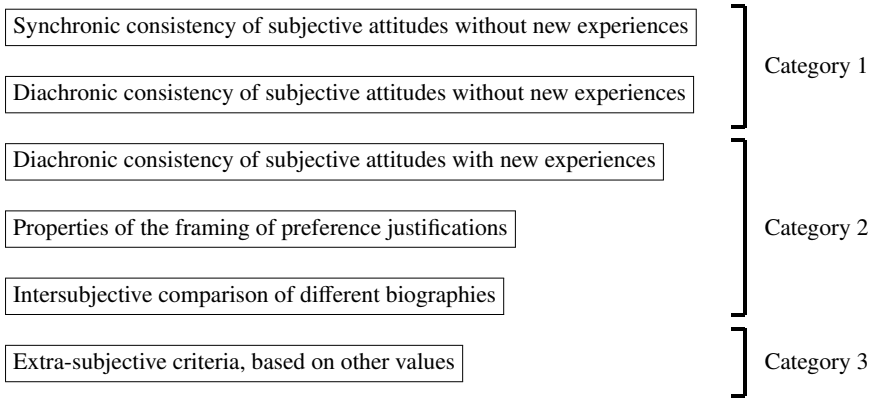

Fig. 11.1

Of these three categories, the third is usually considered to be much more problematic than the other two. Such arguments are considered to be disrespectful, and in cases when their purpose is the person's own good they are seen as paternalistic. In our view, the common reluctance to apply this type of arguments is justified.

[1] Of course, even when employing arguments in the first two categories, the evaluator employs his or her own preferences, namely her (second-order) preferences in favour of preference consistency, preference stability, non-coerced preferences, etc. It is only his or her preferences in the same subject-matter as the contested preferences that we refer to as problematic grounds for discounting the reviewed person's preferences.

Although we do not wish to claim that such arguments should never be used we propose that they should preferably not be used when they can be replaced by arguments belonging to the first two categories. In other words, arguments that consist in reviewing the person's preferences with someone else's preferences in the same subject-matter should only be resorted to when the other two types of arguments cannot be used with the same effect. A procedure for evaluating adaptive (or for other reasons contested) preferences should therefore have two phases. In the first phase, only the first two types of arguments are allowed. Only if the preferences survive the first phase should they be subjected to the second phase, in which arguments in our third category are allowed. It is the purpose of the present paper to outline such a first phase, i.e. to specify in what ways arguments in the first two categories can be used to legitimately discount or dismiss a person's preferences.

11.3 Cases

This section presents a number of exemplary cases, which illustrate the conflicting intuitions about adaptive preferences - some preferences, because they are adaptive, intuitively should be excluded, while others, although they are adaptive, should intuitively not be excluded.

Stockholm Syndrome. Kristin abhorred criminals and their exploits. But when she was taken hostage in a Stockholm bank in the summer of 1973, for five days sharing a vault with her captors and becoming very familiar with them, she wound up emotionally attached to them, showed fear, distrust, and anger toward the authorities, and even defended her captors after the ordeal (de Fabrique et al. 2007).

The preferences that Kristin expressed are adaptive in the sense that she acquired them in response to her being taken captive.[2] We believe that her adapted preferences should not be taken into account when making a decision about how to engage with the criminals. By interpreting her preferences as a reaction to her circumstances, we discount them as a basis for judging her good. Thus, Kristin's case is an example where a reviewee's preferences, because they are adaptive, should be discounted.

Sour Grapes. The fox in Aesop's fable tries to reach some grapes hanging high on the vine but is unable to, although he leaps with all his strength. As he goes away, the fox remarks, "Oh, they aren't even ripe yet! I don't need any sour grapes." (Elster 1985). The fox's preferences are adaptive, because it is only the unavailability of the grapes that makes him change his preferences. It is commonly agreed that the fox's preferences for not eating the grapes should not be considered indicative of

[2] de Fabrique et al. (2007, 14) for example explain the preference change as follows: "Hostages have to concentrate on survival, requiring avoidance of direct, honest reactions to destructive treatment. They must become highly attuned to the pleasure and displeasure reactions of their captors. As a result, victims seem more concerned about the perpetrator's feelings than their own....The hostage endures isolation from other people and has only the captor's perspective to rely on".

his welfare, and that it would be better for the fox to have the grapes. Thus, the fox's case is an example where a reviewee's preferences, because they are adaptive, should be discounted.

Yet the fox's case becomes less clear-cut if we consider his preferences for "not trying to reach the grapes". Here the preferential focus is on the action of reaching rather than the state of having the grapes. The fox might well say to himself "grapes always hang so high that an animal with as short legs as mine would be foolish to try reaching it", while at the same time acknowledging that eating the grapes (in some other circumstances where foxes had longer legs or grapes hung lower) would be desirable. If grapes indeed generally hung out of reach for foxes, then the fox's preference for "not trying to reach the grapes" should not be discounted, despite being adaptive. However, the fox knows, and we know, that this is not generally the case: grapes often hang low enough for even short-legged animals to reach them, thus preferring not to try seems rather an irrational overreaction and hence should be discounted.

Lacking talent. Ron wanted to be an opera singer, and he spent considerable effort and resources on achieving his goal. Yet after some years of professional training and daily long hours of exercise, he realized that he has not progressed much towards his goal, and that he likely lacks the talent to ever get there. Consequently, he stopped his efforts and told his friends "I do not want to become an opera singer anymore, because I realized that I would likely never get there". Like in the fox's case, Ron's preference against becoming an opera singer is clearly adaptive: it is the unavailability of this career goal to him that makes him stop wanting to achieve it. But in contrast to the fox case, we believe that Ron's preference should not be discounted: given his lack of talent, it is not only better for Ron to stop trying to become an opera singer, but it is also better for him not to be an opera singer (belittled by the audience and dismissed by the critics, one can imagine). Thus, Ron's case is an example where a reviewee's preference, despite being adaptive, should not be discounted.

Acquired disability. Due to a genetic condition, Rebecca has gradually become blind over the last 13 years. In the early stages of this process, she was desperate, feeling that loosing her sight impoverished her life. Now, an experimental treatment offers the chance of a cure. Yet now Rebecca says that "would I like to have gene therapy and see perfectly again?...Now I'm not sure" (Atkinson 2007). She argues that "what you lose in one place you gain elsewhere, and while a blind life is different to a sighted life, it is no lesser" (ibid.). Rebecca's preference change seems adaptive: she adjusted her preferences to her new situation, because the old situation was no longer available. Specifically, her assessment of blindness changed from a "loss" to a life that is "different, but no lesser". Yet should we therefore discount her preference? She has, after all, experiences of both a seeing and a blind life, she compares them and finds that her previous strict ranking no longer applies. It thus seems that we should respect her new preference – even if it is adaptive – and not discount it.

Thus, Rebecca's case is an example where a reviewee's preference, despite being adaptive, should not be discounted.

Congenital disability. Nils, in contrast to Rebecca, was born with a disability – a severe hearing impairment. When offered novel treatment (with high but not certain success rate), he spends considerable time considering the pros and cons of the treatment. In the end, he declines. As justification, he cites the vigorous opposition of some members of the Deaf culture to cochlear implants both as a seriously invasive treatment of dubious efficacy and as a threat to Deaf culture. He also cites (hearing) expert opinion that within Deaf culture, there's a level of social intimacy that is rare among the hearing (Dennis 2004, 894). Nils' preference seems adaptive: he formed his preference based on what was available to him, citing other people's opinions about what might be worse in a hearing existence. In his case, unlike Rebecca's, we are less sure whether we should respect his preferences and not search for ways of convincing him to undergo surgery. We treat this as a borderline case, where reasonable persons can have different moral intuitions – which hopefully can be clarified with the analysis proposed in this paper.

Infrared vision. Some day in the near future, Heinrich has received a generous offer. He is a member of a live role-playing community many of whose members have had a "third eye" surgically implanted in their foreheads. Through an advanced interface on the brain surface, the third eye provides rudimentary infrared vision that is said to greatly enhance the experience of role-playing at night. Heinrich has been offered a third eye for free. He spends a lot of time talking to people with long experience of infrared vision, and also to members of the small activist group TIE ("Two Is Enough"). To the surprise of his fellow gamesters, after giving the offer serious consideration he decides to decline it. Would it be right to dismiss or discount his preference for not having a third eye since he does not have the experience of living with one? This case is structurally similar to that of Nils. Probably, most people would be more willing to accept Heinrich's than Nils's preferences, but this difference seems to depend on substantial views on the offered alleged improvements (in these two cases hearing respectively infrared vision). However, such substantial views should not be allowed any influence in the first phase of preference review that we are now focusing on. In a discussion aiming at consistent views on adaptive preferences, reluctance to discount Heinrich's preferences can be used as an argument not to discount Nils's either.

Accepting inferiority. Jayamma is a women in her sixties. She was born in deprived circumstances and worked for four decades in a brick kiln, earning less than the men while doing the same (or even harder) work. Jayamma seemed to have resigned to her situation, accepting discrimination and exploitation (Nussbaum 2001). Her preferences are adaptive because they were formed in response to her circumstances (but see Baber 2007 for criticism of such a "wide" interpretation of adaptive preferences). In that way, Jayammas case is structurally similar to Nils', but unlike in his case, we are convinced that Jayamma's preferences should not be taken as a basis for judging her good. Thus, Jayamma's case is an example where a reviewee's preference, because it is adaptive, should be discounted.

The upshot from considering these different cases is that preferences that are considered adaptive give rise to different normative intuitions: sometimes, we intuit that these preferences should be discounted because of their adaptiveness, yet in other cases, we intuit that they should not be discounted, despite being adaptive. Thus we disagree with a wide array of authors (e.g. Elster 1985, Bruckner 2009, Coburn 2011) who derive an unambiguous normative assessment of adaptive preferences from a uniform analysis of their nature. Instead, we believe that our case collection shows that the normative implications of adaptive preferences require further analysis. In the next section, we propose a formal framework for such an analysis.

11.4 Framework

The essential elements of the framework in which individuals apply their preferences are (i) the objects of preferences and (ii) the background.

The standard representation of objects of preferences is a set of alternatives, whose elements are mutually exclusive. We will assume that there is such an alternative set, denoted $\mathcal{A}$. Its elements represent all the (most detailed, or maximally specified) objects that the agent prefers or disprefers to each other in the context under consideration.

Preferences over such a set of mutually exclusive alternatives have been called exclusionary preferences. Sometimes we express preferences that refer to objects that are not maximally specified, i.e. they refer to combinations of maximal alternatives (combinative preferences). Suppose that we are going to the opera, and there are four shows that we can choose between: *Rigoletto*, *Aida*, *Siegfried* and *Parsifal*. I may then say that I prefer watching a Verdi opera to watching a Wagner opera. This means that I prefer one subset of the alternative set, namely {*Rigoletto*, *Aida*} to another subset, namely {*Siegfried*, *Parsifal*}. A reasonable interpretation of this is that I prefer each element of the former subset to each element of the latter. In another context I may say that I prefer Verdi's operas to Wagner's. Such as statement can be made by someone who prefers most of Verdi's operas to most of Wagner's, but still likes one of Wagner's operas better than one or two of Verdi's. To account for this we can focus on representative pairs of one Verdi opera and one Wagner opera. A person can be said to prefer Verdi's operas to Wagner's if and only if she prefers the Verdi opera in any such pair. (For details, see Hansson 2001.) However, for our present purposes we need not be concerned with such (combinative) preferences. As it will turn out, adaptive preferences can be analyzed with reference to exclusive preferences.

Hence, we can assume that there is an alternative set $\mathcal{A} = \{a_1, a_2 \ldots a_m\}$ whose elements are mutually exclusive, and the preferences we will discuss are preferences over the elements of $\mathcal{A}$. But in addition there is background information, to be denoted $\mathcal{B}$, that has a crucial role. When we say that a person's preferences are adaptive we compare her preference relation over some alternative set $\mathcal{A}$ to some other preference relation that she might have had over that same alternative set. The difference can depend on background information, typically background information

about the plausibility of various conditions that have an influence over the valuation of elements of $\mathcal{A}$. Therefore, the relevant background information to be included in $\mathcal{B}$ should refer to several possible states of affairs and their plausiblity. The latter information can have several different forms: probabilities, probability intervals, plausibility measures, etc. Here, for simplicity, we will make the assumption that they are probabilities. This means that $\mathcal{B}$ can be conceived as consisting of pairs $\langle p_k, \mathfrak{b}_k \rangle$ such that $\mathfrak{b}_k$ is a description of aspects of the world including those that influence the value of various elements of $\mathcal{A}$ (a "background scenario"), and p_k is a probability assignment to $\mathfrak{b}_k$, hence:

$$\mathcal{B} = \{\langle p_1, \mathfrak{b}_1 \rangle, \langle p_2, \mathfrak{b}_2 \rangle, \ldots \langle p_n, \mathfrak{b}_n \rangle\}$$

where we of course assume that

$$\sum_{1 \leq k \leq n} p_k \leq 1$$

We assume that each element of $\mathcal{B}$ contains sufficient information to determine the value of each alternative in $\mathcal{A}$. We will use an expected utility framework, and the desirability of a according to $\mathcal{B}$ is expressed by the following formula:

$$\mathsf{des}_{\mathcal{B}}(a) = \sum_{\langle p_k, \mathfrak{b}_k \rangle \in \mathcal{B}} p_k \times \mathsf{des}_{\mathfrak{b}_k}(a) \tag{11.1}$$

where $\mathsf{des}_{\mathfrak{b}_k}(a)$ is the desirability of a under the circumstances specified by $\mathfrak{b}_k$.

We will call such a pair $\langle \mathcal{A}, \mathcal{B} \rangle$ a *framework* for the person's preferences. In order to evaluate whether preferences are rational (rationalizable) it is useful to search for a framework in which they are rational (rationalizable). If that succeeds then it still remains, of course, to investigate whether the framework itself is rationally defensible.

It could perhaps be argued that the pair $\langle \mathcal{A}, \mathcal{B} \rangle$ is unnecessarily complex. If a person prefers option a_1 to option a_2 under the background condition b, then that can be expressed as a preference for $a_1 \& b$ over $a_2 \& b$ In this way, we can incorporate the background into an augmented alternative set $\mathcal{A}'$ that replaces $\langle \mathcal{A}, \mathcal{B} \rangle$. It is a standard procedure in preference logic to augment the alternative set in this way, if necessary until it consists of the possible worlds, thereby making any background representation redundant (See references in Hansson 2001, 57-60).

Although such an augmented alternative set may be useful for some purposes, when applying preference logic to decision problems we have a strong reason to distinguish between the alternative set and the background. In decision analysis we need to distinguish between that which is under the decider's control and that which is not. The first category contains many of the possible actions of the decider (but not all of them, there are things such as losing weight or quitting an addiction that we cannot be sure to succeed in doing). The second category consists mainly of natural circumstances and acts by others (i.e. of that which is referred to as "states of nature" in decision theory). Usually the first category should be represented as a set

of options that the decider can choose among, whereas the second is more usefully represented as a set of states to which probabilities are assigned. (On this distinction in decision theory, see Rabinowicz 2002.) If we wish to represent preference-based choices or decisions realistically is necessary to keep the two categories apart. A "mixed" object such as $a_1\&b$ in the above example would require in part a probabilistic and in part a non-probabilistic treatment, which is in practice not possible.

11.5 Reconstructibility

Our formal framework gives us a handle for how preferences over alternatives depend on preferences over the partition of states contained in $\mathcal{B}$, in conjunction with the probabilities contained in that background information. In other words, given frame $\langle \mathcal{A}, \mathcal{B} \rangle$ and preferences over $\mathcal{A}$ conditional on states in $\mathcal{B}$, we can reconstruct preferences over $\mathcal{A}$. Let's put this in a formal definition:

Definition 1. *A preference ordering $\succeq$ over elements of $\mathcal{A}$ is* reconstructable *with respect to frame $\langle \mathcal{A}, \mathcal{B} \rangle$ and a desirability function* des *for $\langle \mathcal{A}, \mathcal{B} \rangle$ if and only if it holds for all $a, a' \in \mathcal{A}$ that $a \succeq a'$ if and only if $\mathsf{des}_{\mathcal{B}}(a) \geq \mathsf{des}_{\mathcal{B}}(a')$.*

Reconstructability is a necessary condition for synchronic consistency: if an agent's preferences over $\mathcal{A}$ are consistent, then these preferences are reconstructable with respect to some $\langle \mathcal{A}, \mathcal{B} \rangle$. Note that such preferences might be reconstructable with respect to many different frames $\mathcal{B}$ and desirability functions des, and that not all such constructions are relevant for synchronic consistency. It is however sufficient that the reconstruction represents the agent's actual state of mind.

This poses two potential epistemic problems. First, for an outside observer, it is often difficult to determine which $\mathcal{B}$ an agent believes in. Our model helps the observer in determining those combinations of des and $\mathcal{B}$ that would reconstruct preferences over $\mathcal{A}$, and thus directs investigations that help find out which of these combnations the agent could rationally believe in.

Second, for the preference holder herself, it might not always be clear which $\mathcal{B}$ she believes in. Most of us will know from introspection that our reconstructions of preferences over $\mathcal{A}$ sometimes switch between different $\mathcal{B}$, and that these frame switches between different $\mathcal{B}$ sometimes destabilize our preferences over $\mathcal{A}$. Our model helps the preference holder in determining the space of possible $\mathcal{B}$, as well as determining which of these $\mathcal{B}$ allow reconstructing preferences over $\mathcal{A}$.

Reconstructability with respect to the actual $\mathcal{B}$ is also a sufficient condition for diachronic consistency: if an agent's *change* of preferences over $\mathcal{A}$ are reconstructable with respect to the change from the agent's actual previous $\mathcal{B}'$ to her new $\mathcal{B}$, then that preference change is *prima facie* consistent. We speak of reconstructibility of changes if a preference pattern can be reconstructed as based on a change in the probabilities of background states (changed probabilities of previous elements of $\mathcal{B}$ or added new elements that were not previously there, i.e. they are raised from zero to non-zero probability). Such adjustments are obviously rational provided that the probability changes are rational.

An adaptive preference change can arguably also be rational for other reasons, such as when (i) experience teaches us that something is better or worse than what we thought, or (ii) we have to give up hopes that are psychologically too costly. These factors might be sufficient in their own right for reconstructing a preference change. However, we do not treat these criteria in this paper. They would typically require a change in des. Here we assume that des is constant, in other words if we have $\langle p_k, \mathfrak{b}_k \rangle \in \mathcal{B}$ and also $\langle p'_k, \mathfrak{b}_k \rangle \in \mathcal{B}'$, then the same value $\mathsf{des}_{\mathfrak{b}_k}(a)$ that is used in the valuation of a in $\mathcal{B}$ should also be used in its valuation in $\mathcal{B}'$.

In short: if an agent's assignments of preferences to one and the same alternative set $\mathcal{A}$ differ between two frames $\mathcal{B}$ and $\mathcal{B}'$, then that difference is reconstructible if and only if it can be accounted for in terms of plausible differences in the probabilities of background scenarios assigned by the agent in $\mathcal{B}$ and $\mathcal{B}'$, while keeping the desirability function the same in $\mathcal{B}$ and $\mathcal{B}'$.

11.6 Reconstructing the Cases

In this section we show that each of the examples from section 3 can be reconstructed with respect to either (i) some $\mathcal{B}$ or (ii) some change from $\mathcal{B}$ to $\mathcal{B}'$.[3] Notably, we give possible reconstructions for *all* cases, including those that we intuited should be discounted (Stockholm syndrome, sour grapes, accepting inferiority). Then we show in the next section that our criteria on permissible frames exclude the reconstructions of these cases.

Stockholm Syndrome. In order to reconstruct Kristin's preference change, we take a cue from de Fabrique et al. (2007, 14) explanation that "Hostages have to concentrate on survival, requiring avoidance of direct, honest reactions to destructive treatment. They must become highly attuned to the pleasure and displeasure reactions of their captors."

We can take a_1="aid criminals", a_2="do not aid criminals", $\mathfrak{b}_1$="It is possible to survive without aiding criminals", $\mathfrak{b}_2$="It is not possible to survive without aiding criminals". Assuming for instance that $\mathsf{des}_{b_1}(a_1) = 90$, $\mathsf{des}_{\mathfrak{b}_1}(a_2) = 110$, $\mathsf{des}_{\mathfrak{b}_2}(a_1) = 100$, $\mathsf{des}_{\mathfrak{b}_2}(a_2) = 0$, it is easy to show that a_1 is preferred to a_2 if the probability of $\mathfrak{b}_2$ is high enough. Hence we can reconstruct her preference change with respect to frame $\langle \mathcal{A}, \mathcal{B}' \rangle$.

Sour Grapes. Before trying to reach the grapes, the fox prefers eating the grapes (a_1) to not eating them (a_2). He considers four background states, $(\neg)\mathfrak{r}\&(\neg)\mathfrak{s}$ where $\mathfrak{r}$ says that the grapes are sour and $\mathfrak{s}$ that an attempt to get hold of some of them will be successful. (Note that the probability of $\mathfrak{r}$ is always above zero, as grapes may fall down.) These are the four elements of $\mathcal{B}$. The desirability of eating the grapes will be high in the case $\neg\mathfrak{s}$, low in the case $\mathfrak{s}$, and independent on whether $\mathfrak{r}$ or $\neg\mathfrak{r}$ pertains. Now when the fox learns that he cannot reach the grapes, this will have

[3] For some cases, this leads us also to reconsider the set $\mathcal{A}$, thus yielding possible alternative reconstructions.

no effect on the probabilistic relation between the a_i and $\mathfrak{s}$ or $\neg\mathfrak{s}$. Hence we cannot reconstruct his preference change with respect to frame $\langle \mathcal{A}, \mathcal{B} \rangle$.

However, the situation looks different if the fox considers preferences over the alternatives "trying to get hold of the grapes" (a'_1) or "not trying to get hold of them" (a'_2). (Note that by changing $\mathcal{A}$ to $\mathcal{A}'$, we open a whole new reconstruction project). These alternatives depend on all four background states $(\neg)\mathfrak{r}\&(\neg)\mathfrak{s}$. The desirability of trying to eat the grapes will be high in the case $\mathfrak{s}\&\mathfrak{r}$, low in the other cases. The desirability of *not* trying to eat the grapes, in contrast, will be low in the case $\mathfrak{s}\&\mathfrak{r}$, high in the other cases. Say, for example, $\mathcal{B} = \{\langle 0.4, \mathfrak{r}\&\mathfrak{s}\rangle, \langle 0.05, \mathfrak{r}\&\neg\mathfrak{s}\rangle, \langle 0.5, \neg\mathfrak{r}\&\mathfrak{s}\rangle, \langle 0.05, \neg\mathfrak{r}\&\neg\mathfrak{s}\rangle\}$. Because the probabilities of $\mathfrak{r}\&\neg\mathfrak{s}$ and $\neg\mathfrak{r}\&\neg\mathfrak{b}s$ are small, we may neglect them here. Consequently, for $\mathsf{des}_{\mathcal{B}}(a'_2) > \mathsf{des}_{\mathcal{B}}(a'_1)$, it must be the case that $1.25[\mathsf{des}_{\neg\mathfrak{r}\&\mathfrak{s}}(a'_2) - \mathsf{des}_{\neg\mathfrak{r}\&\mathfrak{s}}(a'_1)] > [\mathsf{des}_{\mathfrak{r}\&\mathfrak{s}}(a'_1) - \mathsf{des}_{\mathfrak{r}\&\mathfrak{s}}(a'_2)]$. This does not seem plausible, as it requires that the benefit from reaching sweet grapes is smaller than the disadvantage from trying and failing to get them, multiplied by factor 1.25. The fox wold have rather weak tastes for sweet grapes if that was the case. But clearly, as $p_{\mathfrak{r}\&\mathfrak{s}}$ goes down, for most people's (or foxes') tastes, there will be a threshold where it is perfectly reasonable to prefer a'_2 to a'_1. $\mathcal{B}' = \{\langle 0.05, \mathfrak{r}\&\mathfrak{s}\rangle, \langle 0.05, \mathfrak{r}\&\neg\mathfrak{s}\rangle, \langle 0.85, \neg\mathfrak{r}\&\mathfrak{s}\rangle, \langle 0.05, \neg\mathfrak{r}\&\neg\mathfrak{s}\rangle\}$, we believe, would be such a fame for us. Hence we can reconstruct his preference change with respect to some frame $\langle \mathcal{A}', \mathcal{B} \rangle$ with relatively high $p_{\mathfrak{r}\&\mathfrak{s}}$, but not with respect to another $\langle \mathcal{A}', \mathcal{B}' \rangle$ with sufficiently low $p_{\mathfrak{r}\&\mathfrak{s}}$.

Lacking talent. Ron has preferences over alternatives of becoming an opera singer (a_1) or not (a_2). The relevant background has two states, talent and not talent. The desirability of becoming an opera singer will be different in these two background states, and the composite desirability will depend on the probabilities of the two background states. This can be reconstructed in terms of changes from frame $\langle \mathcal{A}, \mathcal{B} \rangle$ in which the probability that he has the desired talent is high to a frame $\langle \mathcal{A}, \mathcal{B}' \rangle$ in which it is low.

Acquired disability. Before the onset of her disease, Rebecca prefers having sight (a_1) to being blind (a_2). The states in her background conditions are two, "blindness impoverishes sensory experience" ($\mathfrak{b}_1$), and 'blindness does not impoverish sensory experience" ($\mathfrak{b}_2$). Before the onset of her disease, her preference for a_1 to a_2 was reconstructible from background beliefs $\mathcal{B}$, according to which the probability of $\mathfrak{b}_2$ was low. Yet during her gradual loss of sight, she learns that this is not true: other senses become sharpened as her sight fades, thus increasing the probability of $\mathfrak{b}_2$. Hence we can reconstruct her preference change with respect to a new frame $\langle \mathcal{A}, \mathcal{B}' \rangle$.

Congenital disability. Nils prefers being deaf (a_1) to having hearing (a_2). The states in his background conditions are two, "Deaf culture offers a more intimate social network than majority culture" ($\mathfrak{b}_1$), and this not being so ($\mathfrak{b}_2$). Because Nils believes that the kind of intimacy he craves can only be found in Deaf groups, he considers

the probability of ($\mathfrak{b}_2$) to be low. Hence we can reconstruct his preferences with respect to a frame $\langle \mathcal{A}, \mathcal{B} \rangle$, in which the probability of $\mathfrak{b}_2$ is very low.

Infrared vision. This case can be reconstructed in the same way as that of congenital disability. Just let a_1 denote not having the infrared vision implant ("third eye") and a_2 having it. Let $\mathfrak{b}_1$ denote "Infrared vision will improve my experiences in role-playing and other situations that are important to me", and let $\mathfrak{b}_2$ denote that this is not so.

Accepting inferiority. Jayamma prefers having less say and a more restricted life than men (a_1) to fighting against such conditions (a_2). The relevant background states in this case, presumably, include "women who demand equal rights as men can secure minimal needs and being part of a community" ($\mathfrak{b}_1$), and "such women cannot secure minimal needs and being part of a community' ($\mathfrak{b}_2$). As far as we interpret the example, Jayamma believes the probability of $\mathfrak{b}_1$ to be very low. Hence we can reconstruct her preferences with respect to a frame $\langle \mathcal{A}, \mathcal{B} \rangle$, in which that is the case. (See Baber 2007 for a similar interpretation that sees Jayamma and related women as "behaving prudently" given her beliefs about background conditions.)

It is notable that in all these cases, the preferences or preference changes of each case could be reconstructed. This contradicts our intuitions in the cases of Stockholm syndrome, Sour grapes, Accepting inferiority and possibly Congenital disability.

Two comments might help analyze this contradiction. One is to stress that we only operated with hypothetical frames. When investigating people's frames $\langle \mathcal{A}, \mathcal{B} \rangle$ empirically instead, it might well transpire that the thus identified frames do not allow the reconstruction of certain preferences. An empirically enriched analysis could thus remain within the confines of consistency and still make substantial claims on which preferences should be discounted or excluded. (For a proposal in this direction, see Bovens 1992).

Another way is to impose constraints on the kind of frames that are normatively admissible for the reconstruction of preferences. This is important both for the many cases where an empirical analysis is difficult and costly, and in particular for those cases where the empirical investigation yields a person's actual frame that makes her preferences reconstructible, even though intuitively it should be discounted. The Jayamma case is salient here: even if Jayamma really held this frame, and thus her preference for discrimination and exploitation were reconstructible, we would not want to accept her preference as identifying her good. Thus, in the next section, we develop some criteria for the admissibility of frames.

11.7 Admissible and Inadmissible Reconstructions

As we showed in the last section, all our cases – even the counterintuitive ones – can be reconstructed with some frame $\langle \mathcal{A}, \mathcal{B} \rangle$ and some desirability function des. This seems too bad for our criterion. Of course, one could argue that a more detailed investigation into each of these cases will reveal what the actual $\mathcal{B}$ of the agents in

each case is, and that this would likely result in some of these cases being excluded on consistency grounds. But a lingering concern remains: would we really say that in all cases, if preferences over $\mathcal{A}$ are reconstructable from an agent's actual $\mathcal{B}$ and some des, then those preferences cannot be excluded on procedural grounds? In particular, think of the Jayamma case: she may indeed hold background beliefs and a desirability function that reconstruct her preferences for inferior treatment of women and in particular a strongly privileged position of her husband in comparison to her. Is a procedural account unable to exclude such preferences from being relevant for a welfarist account of the good?

We think not. Rather, we offer a set of criteria for judging which frames are admissible for reconstruction. Criticism of a person's preferences (and the ensuing discounting of them) can refer either to the preference framework $\langle \mathcal{A}, \mathcal{B} \rangle$ or to the value assignments that the person makes in that framework. To begin with the first element of $\langle \mathcal{A}, \mathcal{B} \rangle$, an adequate alternative set should consist of exactly those relevant options that are within the person's control. Two major types of criticism are possible here:

(i) Some element of $\mathcal{A}$ is not really under the reviewed person's control.
Example: The reviewed person expresses preference for the option of using addictive drugs in the next six months, and then stop using them.

(ii) Some relevant option that is under the reviewed person's control is not included in $\mathcal{A}$.
Example: A battered wife believes that if she leaves her husband she will not be able to refrain from going back to him, in spite of evidence showing that others in a similar situation have been able to do this.

Criticism concerning $\mathcal{B}$ would have to refer to its realism:

(iii) $\mathcal{B}$ is not within the scope of reasonably realistic views of what the world is like.
Example: A battered wife believes that if she serves her husband better food then he will stop beating her.

Finally, concerning the preference relation itself we propose that there are two forms of criticism that do not transgress the limitation that we set in Section 2, namely that it should not refer to other preferences over $\mathcal{A}$ than those of the reviewed person: the preferences may have been formed under conditions with either external or internal limitations on the preference formation process.

(iv) The preferences or the framework have been formed under external circumstances that prevent the free formation of preferences, such as coercion or information restrictions.
Example: A person who has been brought up in a sect, and does not know the outside world, develops a preference for becoming a suicide bomber.

(v) The reviewed person lacks the mental capacities necessary to form independent preferences.
Example: A mentally handicapped person decides to give away her savings to a nurse who has been kind to her.

Focussing on the causal influence on the frame, rather than the preference itself, we thus address the relevance of adaptation for the justification of these preferences. As Bruckner (2011) recently pointed out, many accounts of adaptive preferences wrongly take the compromising circumstances of the formation of preferences as necessarily compromising the circumstances of the agent's justification of her preferences. This dependence is not necessary: a preference that I acquired through in some dubious fashion (e.g. through some kind of adaptation) might still be normatively relevant if I have legitimate reasons for holding it. Our approach avoids this confusion. It shows when and how adaptation affects preferences through affecting their (justificatory) background, and through that identifies those cases of adaptation that indeed compromise a preference's justification.

Let us now consider the extent to which these five types of preference criticism can be applied to our six examples of adaptive preferences, as reconstructed in Section 6. In two cases, namely Ron (the opera student) and Rebecca (who was offered a cure against her blindness) none of the criteria (i)-(vi) seems to be applicable. This means we have admissible reconstructions with the help of which the adaptive preferences in these two cases have successfully passed the first phase of the evaluation.

The cases of Nils (who was offered a treatment against congenital deafness) and Heinrich (who was offered a "third eye" imparting infrared vision) are more doubtful. It could be argued that their preferences have been formed under conditions that violate the part of criterion (iv) that refers to information restrictions. Niels has always been deaf and Heinrich has always been unable to see infrared light (or "infrared blind" as some of his three-eyed friends would say). Therefore, their information on what it means to lack these sensory faculties can only be indirect and not based on personal experience. However, since both Nils and Heinrich have had access to, and have seriously considered, the experiences of others with these faculties, we propose that criterion (iv) cannot be used to dismiss or discount their respective preferences. The contrary standpoint would put us on a slippery slope, and anyone's preferences concerning something that others have more direct knowledge of than they have themselves might have to be discounted. Our tentative conclusion is therefore that Nils's and Heinrich's preferences should be treated in the same way as Rebecca's, i.e. the reconstruction should be accepted, and their adaptive preferences have successfully passed the first phase of the evaluation.

The case of Kristin (the hostage exhibiting the Stockholm syndrome) is much simpler. Her preferences were formed under coercive conditions that starkly violate criterion (iv). As de Fabrique et al. (2007, 14) point out, "the hostage endures isolation from other people and has only the captor's perspective to rely on". This failure concerns her preferences as such, not only a particular reconstruction such as that proposed in Section 6. Therefore the failure cannot be mended with any type of reconstruction, and we conclude that her adaptive preferences *cannot* successfully pass the first phase of the evaluation.

Turning to Jayamma (who prefers a life in subordination and exploitation), the reconstruction presented in Section 6 violates condition (ii), since the set of alternatives does not include options that are in fact available to her, such as leaving her

present community and starting a new life elsewhere. Condition (iii) also seems to be violated since her background beliefs are unrealistic with respect to what is possible for her. There is ample of evidence for why such a misspecification can happen and what the roe of the respective institutions in this are:

"For the vast majority of Indian women, these statutes [equal rights, including economic independence, for women; the right to education; the right to free exercise of religion], are meaningless...Lack of basic knowledge about the law and procedures, delays and insensitivity of the judicial system, the cost involved in getting justice have all contributed to this." (Nussbaum 1999, 398, citing report by United Nations Population Fund)

From this it can be concluded that this reconstruction of her adaptive preferences has failed in the first phase of the evaluation. However, since this failure concerns this particular reconstruction and not her preferences as such, we cannot conclude that no other reconstruction can be more successful.

11.8 Conclusion and Outlook

We have proposed the outlines of a general structure for the evaluation ("stress-testing") of adaptive preferences. It consists of two phases. In the first of these the reviewer is not allowed to appeal to her own values in the same subject matter, but in the second phase such appeals are allowed. The reason for this two-phase structure is that the imposition of one's own values is problematic and should therefore be avoided as far as possible.

We have also shown how the first phase can be constructed. The construction is based on the two assumptions (1) that rational (adaptive) preferences should be reconstructible in terms of a precise framework consisting of an alternative set and a set of background beliefs, and (2) that such a reconstruction can be rejected if the alternative set or the set of background beliefs misrepresents the actual conditions or if the preferences have been formed under conditions with external or internal limitations that distort the preference formation process. Our examples demonstrate how adaptive preferences can be reconstructed in the sense just referred to and how the evaluation of such reconstructions lead, in different cases, to either positive or negative conclusions on the adaptive preferences in question.

In order to make this evaluation structure complete, more details on this first phase need to be filled in. In addition, the second phase needs to be structured and criteria have to be developed that distinguish between admissible and non-admissible applications of one's own preferences when evaluating the preferences of someone else in the same subject matter.

Acknowledgements. Till Grüne-Yanoff and Sven Ove Hansson gratefully acknowledge financial support from *Riksbankens Jubileumsfond* through the project *Prioritizing in Health Care*.

References

Atkinson, R.: Do I Want My Sight Back? The Guardian, July 17 (2007)
Baber, H.E.: Adaptive Preference. Soc. Theory Pract. 33(1), 105–126 (2007)
Bovens, L.: Sour Grapes and Character Planning. J. Philos. 89(2), 57–78 (1992)
Bruckner, D.W.: In Defense of Adaptive Preferences. Philos. Stud. 142(3), 307–324 (2009)
Bruckner, D.W.: Colburn on Covert Influences. Utilitas 23(4), 451–457 (2011)
Colburn, B.: Autonomy and Adaptive Preferences. Utilitas 23(1), 52–71 (2011)
de Fabrique, N., Romano, S.J., Vecchi, G.M., van Hasselt, V.B.: Understanding Stockholm Syndrome. FBI LEB 76(7), 10–15 (2007)
Dennis, C.: Genetics: Deaf by Design. Nature 431(7011), 894–896 (2004)
Elster, J.: Sour Grapes: Studies in the Subversion of Rationality. Cambridge University Press, Cambridge (1985)
Hansson, S.O.: The Structure of Values and Norms. Cambridge University Press, Cambridge (2001)
Mazie, S.V.: Consenting Adults? Amish Rumspringa and the Quandary of Exit in Liberalism. Perspect. Polit. 3(4), 745–759 (2005)
Nussbaum, M.C.: In Defense of Universal Values. Idaho L. Rev. 36, 379–443 (1999)
Nussbaum, M.C.: Symposium on Amartya Sen's Philosophy: 5 Adaptive Preferences and Women's Options. Econ. Philos. 17(01), 67–88 (2001)
Rabinowicz, W.: Does Practical Deliberation Crowd out Self-Prediction? Erkenntnis 57, 91–122 (2002)

Chapter 12
Refusing Life-Saving Treatment, Adaptive Preferences, and Autonomy

Jukka Varelius

12.1 Introduction

Consider a case of a patient receiving life-supporting treatment. With appropriate care the patient could be kept alive for several years. Yet his latest prognosis also indicates that his mental abilities will deteriorate significantly and that ultimately he will become incapable of understanding what happens around and to him. Despite his illness, the patient has been eager to live. However, he finds the prospect that he is now faced with devastating. He undergoes an unconscious process that results in his finding himself with a preference to die. Consequently, he requests his doctor to stop the treatment that she is providing. Other things being equal, standard medical ethics and the legislations of most Western countries allow the doctor to adhere to the patient's request if it is autonomous.[1] While its origins can be traced back to at least J.S. Mill's work on women's rights,

Jukka Varelius
Department of Philosophy and Behavioural Sciences
University of Turku
Turku
Finland
e-mail: jukvar@utu.fi

[1] While no universal consensus exists on the precise definition of (voluntary) euthanasia (see, e.g., Beauchamp and Davidson 1979), stopping the life-supporting treatment of a patient at his autonomous request is commonly called passive voluntary euthanasia. The moral status of active voluntary euthanasia – autonomously requested euthanasia in which a doctor has a more involved role in bringing about her patient's death than merely that of withholding or withdrawing life-supporting treatment from him – remains more controversial and the practice is prohibited by law in most countries. For this reason, my main example is a case of passive voluntary euthanasia. Yet the considerations I present also apply in the case of active voluntary euthanasia.

J. Räikkä and J. Varelius (Eds.): *Adaptation and Autonomy*, SAPERE 10, pp. 183–197.
DOI: 10.1007/978-3-642-38376-2_12 © Springer-Verlag Berlin Heidelberg 2013

unconscious alteration of one's preferences in light of the options available has figured, for example, in recent discussion on female oppression and utilitarian social choice theory (see, e.g., Elster 1985, Mill 1988, Nussbaum 2001, Sen 1995). Yet the topic has not received attention in the debate on voluntary euthanasia.[2] The inattention is unfortunate because the central problem with such adaptation has been taken to be that it undermines personal autonomy. This motivates the main question examined in this chapter, *i.e.* the problem of whether a request for euthanasia based on an adaptive preference is autonomous.

12.2 Clarifying the Main Concepts and the Main Problem

The phenomenon focused on here can be elucidated in terms of the familiar fable, *The Fox and the Grapes*. In the story, a hungry fox tries to reach some grapes hanging high on the vine. After realizing that it cannot get to them, it turns away from the grapes and declares that they would be too sour for its tastes anyway. In his influential discussion on the phenomenon, Elster (1985, 25) describes preference adaptation as follows:

> *Adaptive preference formation* is the adjustment of wants to possibilities – not the deliberate adaptation favoured by character planners, but a causal process occurring non-consciously. Behind this adaptation there is the drive to reduce the tension or frustration one feels in having wants that one cannot possibly satisfy.

Hence, in acquiring an adaptive preference a person molds his preference in light of the options he perceives himself as having. This happens unconsciously; the person is not aware of what he is doing. An adaptive preference as it is here conceived is thus one that is not arrived at as a result of rational reflection, by consciously assessing what one has most reason to prefer given the facts of the situation one faces. Instead, it is a preference one just finds oneself with after realizing that one cannot get what one wants the most; adaptive preference change takes place behind one's back, so to speak (Elster 1985, 117). Though adaptive preference change can refer merely to the acquisition of an adaptive preference, gaining such a preference is always a part of a larger process. The process

[2] Speaking of philosophical literature more inclusively, Bruckner (2009, 308) says that since the publication of Elster's (1985) "groundbreaking work, there has been an inadequate and unsteady trickle of literature devoted to adaptive preferences." A recent exception to the rule that preference adaptation has not received attention from medical ethicists is Courtwright (2009, 93), who refers to the phenomenon in arguing that stigmatization can lead to health inequalities. According to him, one way in which that can happen is that unconscious internalization of stigmatization causes members of the stigmatized groups to have and demand less income, education, and other social determinants of health. The notion of adaptation has also figured in the discussion on the so-called 'disability paradox' (see Hanna's and Schermer's contributions to this volume).

involves an agent with a certain initial want, her perception that she is unable to satisfy it, a psychological drive to alleviate the disappointment and lessen the tension the perception causes, and the unconscious change of preference resulting from the functioning of the drive.[3]

Of course, what someone deems an adaptive preference may not be an adaptive preference and *vice versa*; mistakes can obviously happen in this context too. It is also plausible that not all requests for euthanasia are based on adaptive preferences. Yet I assume that some of them are and, thus, that assessing whether the request the patient in the above example presents is autonomous is not practically futile on the grounds that such cases will not be encountered in the real world. The central concern that Elster (1985, 125 ff.) has with adaptive preference change relates to the effect of the phenomenon on self-government; in his view, the unconscious alteration of preferences is not autonomous. In a similar vein, Colburn (2011, 65-69, cf. Bruckner 2011), for another example, maintains that preferences that have what he calls covert explanations – explanations of which the ones having the preferences are necessarily unaware – are heteronomous and that adaptive preferences are paradigm examples of preferences that have such explanations.[4]

Beauchamp and Childress explicate the concept of personal autonomy that figures centrally in contemporary biomedical ethics in terms of the notions of intentionality, understanding, and freedom. According to the authors, an autonomous person acts freely and intentionally in accordance with her own informed view on how her life should proceed. For an act to be intentional in the relevant sense it must correspond to the agent's conception of the act in question; an agent who acts intentionally does what she thinks she does. In terms of the kind of understanding autonomy presupposes, Beauchamp and Childress write that, at a minimum, persons understand only if they have acquired pertinent information and have relevant beliefs about the nature and consequences of their actions. Generally autonomy does not presuppose complete understanding, they continue, but, in some cases, a person's lack of awareness of even a single risk or missing fact can deprive her of adequate understanding. A person lacks freedom in the relevant sense when she is under the influence either of external sources or

[3] As well as this phenomenon, 'adaptation' has also been used to refer to the changes that organisms or species undergo to accommodate themselves to particular environments, functional components of the nervous system that solve particular reproductive problems, the process in which a continuous presence of a certain stimulus makes receptors less sensitive to the stimulus, etc. These kinds of senses given to the term are now put aside; here adaptation refers solely to unconsciously altering one's preferences in light of available options.

[4] Not all authors writing on preference adaptation think that the alteration of preference must be unconscious (see, e.g., Bovens (1992); c.f., e.g., Zimmerman (2003)). However, when the preference change is consciously undertaken, the question would rather be about what Elster (1985, 25) calls character planning. And it appears unclear why it would be problematic from the viewpoint of the patient's autonomy if his request for euthanasia was made consciously (see also below).

internal states that rob her of self-directedness. The main examples of such external influences are coercion and manipulation by others. Phobias and compulsions are paradigm examples of internal states that deprive one of autonomy (see Beauchamp and Childress 2009, 99 ff., Beauchamp 2005, 2010, and also, e.g., Secker 1999).[5]

While preference adaptation has thus been considered problematic from the viewpoint of autonomy,[6] adjusting one's preferences in light of what is possible would also appear eminently reasonable. It might indeed be claimed that we should rather worry about the autonomy of those people whose preferences do not adjust to the realities of their lives than of the self-government of those whose preferences do adapt. For example, commenting on Elster's views on the autonomy of adaptive preferences, Sandven (1999b, 175) writes as follows:

> There would rather be reason to question the person's autonomy if he [a person who is unable to satisfy his want(s)] continued indefinitely to long for something he should clearly have understood was unattainable for him. … Rather than regarding such insensitivity in the "preferences" as a characteristic of autonomy and start looking for explanations in cases where they are changed when what one desires turns out to be unattainable, we may more fruitfully turn the question on its head and ask why one should persist in desiring what has turned out not to be in harmony with one's capabilities and potentials.

---

[5] To clarify, this conception of autonomy does not entail that a person can be autonomous only if she always achieves her goals and never diverts from her own conception of how her life is to proceed. It also does not presuppose that the self and the goals of an autonomous person cannot be affected by her environment or by the persons around her. For discussion on these topics see, e.g., Christman (2009) and May (2005). Moreover, autonomy as it is conceived here thus allows for degrees. Accordingly, while a person who has complete understanding is more autonomous than one who has not, even the latter can be able to make an autonomous choice as long as she has sufficient understanding. The main argument of this article does not presuppose determining precisely what degree of autonomy a patient must possess to be able to make an autonomous choice about ending his life. Consequently, I now put this question aside and speak only about sufficient understanding. Finally, the authors whose views are discussed below do not necessarily understand autonomy in precisely the same sense as the one adopted in this article. As my purpose is not to criticize their views, but to assess whether the grounds for deeming preference adaptation autonomous they propose show that such preference alteration can be autonomous in the sense focused on now, whether they do does not affect the main argument of this chapter. Although the sense of autonomy adopted here is not universally endorsed, even those who want to understand autonomy differently or prefer not to use the term 'autonomy' at all may find value in actions that are intentional, informed, and free in the senses just briefly characterized.

[6] Under a different name, a classic defense of the value of autonomy is found in J.S. Mill (1985). Depending on the interpretation, Mill saw that value either as instrumental to wellbeing, constitutive of wellbeing in a wide sense, or intrinsic. For more recent arguments in support of the significance of autonomy see, e.g., Haworth (1986), Harris (2003), and Kraut (2007).

Adopting this view, it might be taken that the fact that the above patient's request for euthanasia is based on an adaptive preference is not problematic from the viewpoint of his autonomy at all. What then should one think here: is the request heteronomous because it is based on an adaptive preference or not? To the extent that the above characterization of the central concepts and the main example of this article are not sufficient for examining this problem, they will be specified in pertinent connections below.[7]

12.3 How Does Preference Adaptation Affect the Autonomy of the Patient's Choice?

In extreme cases, undergoing a traumatic event can lead to psychosis (see, e.g., Morrison, Frame and Larkin 2003). Serious deeds that violate their victims' bodily and mental integrity – such as acts of severe violence and sexual abuse – are key examples of such events. Yet ordeals with comparably grave consequences can plausibly also happen in other contexts in which people have highly important interests at stake. While reactions to being given a diagnosis of a serious illness are known to vary, receiving bleak news about her health can sometimes shatter a person quite profoundly (cf., e.g., Ford, Lewis and Fallowfield 1995; and Block 2001). Consider furthermore the phenomenon commonly referred to as self-deception. Though the precise way in which this happens remains obscure and controversial (see, e.g., Borge 2003; Levy 2004; and Mele 2009), there are cases in which people cling to a belief despite being provided with overwhelming evidence of its falsity. The examples include the desperate drug addict who is positive that her dependence is totally under her control, the self-indulgent person

[7] As I focus on autonomy understood in the procedural sense briefly described above, I bypass former attempts to solve perceived problems related to preference adaptation that presuppose substantive or Kantian conceptions of autonomy. Perhaps many autonomous people value quite similar things. For example, Nussbaum's (2001, 83, 87-88) – who understands adaptation in a somewhat different sense than the one adopted here (see Räikkä's contribution to this volume) – list of the central human capabilities may indeed qualify, as she purports it, as a partial account of the human good sufficient for a basis of political principles. And as consulting all concerned parties can be impossible, or at least quite inefficient, in connection with deciding how to arrange or develop societal institutions, it appears quite reasonable to refer to some substantive and assumedly objective notion of human good in connection with making those kinds of choices. But the view that there is a precisely definable list of substantive goods that all autonomous people necessarily value seems rather difficult to justify (cf., e.g., Christman 2009). Accordingly, at least as things are, requiring that the treatment choices individual autonomous patients make in their own matters must conform to some such substantive starting point appears unwarranted. For an argument to the effect that referring to the notion of autonomy adopted here does not itself lead to unjustifiably (dis)favoring particular lifestyles see, e.g., Colburn (2007).

who refuses to believe that his unhealthy lifestyle is affecting his condition, and the jealous wife who is convinced of the infidelity of her devoted husband.

When the disparity between how things are and how the individuals would want[8] them to be is great and the question is about matters they find significant, the persons can be unable to accept the facts they are faced with. Consequently, they can remain in a state of self-deception for quite considerable periods despite the possibility of serious consequences – such as loss of self-respect, health, and life – that can have for them. The mental forces operative in these kinds of cases are thus quite potent and can have even drastic effects on personal autonomy. Similarly, it would seem that the disappointment and frustration caused by a patient's inability to satisfy his wants can make him unable to come to terms with the facts he faces. That is most likely to happen when the wants which the patient cannot satisfy are important to him, as they usually are when patients who are eager to live are told that living the kind of life they consider worth living will cease to be an option for them. In such cases, the patients' autonomy would appear to be in danger. Considering our main example in terms of the three elements of autonomy distinguished above supports this position.

12.3.1 Adaptive Preference Change and Intentionality

A patient's request for euthanasia is intentional in the above-described sense when what he does in presenting it corresponds to his conception of the action in question, *i.e.* that he is requesting euthanasia. Accordingly, the plea that his treatment be stopped the patient in our main example presents amounts to an intentional euthanasia request if the patient understands that withdrawing the treatment that he is receiving results in his death. If the psychological drive to reduce the tension that getting the devastating prognosis causes is strong enough to confuse the patient so that he does not realize that cessation of his treatment would have such a consequence, his call for stopping his treatment does not amount to an intentional request for euthanasia. Whether a patient has the kind of comprehension that intentionally requesting for euthanasia presupposes and whether his euthanasia request actually is intentional are ultimately empirical questions. Thus, preference adaptation may, but need not, undermine a patient's ability to intentionally request euthanasia; whether it does so must be assessed on each, individual case.

12.3.2 Adaptive Preference Change and Understanding

In philosophical literature on adaptive preference change *The Fox and the Grapes* fable has been interpreted in two significantly different ways. First, it has been taken that the fox acquires a false belief about the taste of (the) grapes – that they

[8] In the case of the overly jealous wife the question is rather about a fear than a want. Her case is thus an example of what has been called twisted self-deception, self-deception that involves the acquisition of an unwanted belief.

are too sour for its tastes when in fact they are not (see, e.g., Berges 2011, 73) – and changes its preference accordingly. Second, it has been taken that the fox changes its preference toward (the) grapes without altering its beliefs about their flavor (cf., e.g., Bovens 1992). There are corresponding two variations of our main example. After realizing that continuing the kind of existence he has deemed worthwhile ceases to be an option for him, the patient acquires new beliefs about life. In light of these new beliefs, life is not worthwhile after all and his preference toward living changes accordingly. Alternatively, the patient's beliefs about life remain as they were, only his attitude towards living changes. Both of these two variations of the case appear problematic in terms of the kind of understanding that autonomy presupposes.

The new beliefs the patient acquires in the first interpretation of the case are, *ex hypothesi*, not based on his assessing the relevant facts of the matter. Instead, the cause of his attaining the new beliefs is that he acknowledges that continuing the kind of life he has considered worthwhile is becoming impossible for him. The patient's new negative beliefs about life may, but need not, accurately depict the kind of existence he could still have. But, because his acquisition of those new beliefs is not based on his considering the facts of the matter, even if they correctly described the kind of life he could still have, that would be merely accidental. Thus, the patient is not justified in having the new beliefs about life on the basis of which he ends up requesting euthanasia.[9] Accordingly, the patient lacks sufficient understanding of the grounds of his presenting the request for it to be autonomous.

In the second version of the case, the patient thus retains the beliefs about life that were the basis of his former assumedly autonomous judgment that life is worth living and only his preference towards life changes. Yet it appears that the preference alteration could be based on sufficient understanding only if the patient now had something on which to base it. What reason would he have to alter his preference towards life when his beliefs about life – on the basis of which he formerly made the autonomous judgment that life is worth living – remain exactly as they were? That he could reasonably conclude that he was mistaken about the value of life – and would thus need to change his views because of that – would seem to presuppose that he has acquired some new beliefs about life. But, *ex hypothesi*, in this second version of the case that does not happen. The beliefs about life he has could, of course, allow for different kinds of judgments about its value; in light of them, his glass could be either half full or half empty. Yet, even if that were the case, the current alteration of his conception of the value of life would still seem to require some grounds to be reasonable. Why would his glass now be half empty when formerly it was half full? Accordingly, in both of the two interpretations of the case, the patient would appear to lack the kind of understanding that making an autonomous request for euthanasia presupposes.

[9] This is not to say that having adequate understanding and possessing a justified belief always presupposed considering the facts relevant to the truth of the belief. But in cases in which the question is about making as important decision as that of whether or not to continue living, the requirement appears appropriate. I return to this issue below.

12.3.3 Adaptive Preference Change and Voluntariness

Some authors maintain that only intentional control by others can undermine the voluntariness of a patient's choice (cf., e.g., Appelbaum, Lidz, and Klitzman 2009).[10] Yet it seems clear that a phobia or a compulsion can undermine the voluntariness of a patient's choice though, or even when, they are not intentionally controllable by anyone. Accordingly, I take it that the possibility that preference adaptation can undermine voluntariness even when it does not result from intentional action by others should not be ruled out to begin with. And it indeed seems that adaptive preference change does preclude voluntary choice even when it is not intentionally induced by others. Because of the influence of the mental drive to reduce frustration operative in preference adaptation, a person can be led to acquire unwarranted beliefs about matters relevant to his choice and to alter her preferences without having sufficient grounds for doing so. Hence, what happens in connection with adaptive preference change is reminiscent of manipulation in terms of its effect on the person's self-directedness.[11] Both a manipulated person and a non-manipulated person who chooses on the basis of an adaptive preference are led – by the manipulator(s) and by a mental drive to reduce frustration respectively – to choose in light of insufficient understanding of the grounds of their choices. Accordingly, like successful manipulation, also preference adaptation results in an involuntary choice. Assessing how preference adaptation affects the patient's ability to make an intentional, informed, and voluntary request for euthanasia thus buttresses the view that his adaptive preference is not autonomous.[12]

---

[10] Appelbaum, Lidz, and Klitzman (2009) also maintain that only control that is not warranted by commonly acceptable moral norms can undermine voluntariness. For criticism of such a moralized notion of voluntariness see, e.g., Olsaretti (1998, 2008).

[11] When adaptive preferences are intentionally induced by others (cf. Courtwright 2009) the question is (arguably) about a form of manipulation.

[12] Sandven (1999a) argues that preference adaptation can be autonomous even if it is not of the intentional kind that Elster calls character planning. According to Sandven, the process of mourning within which one adapts to a loss is usually not best described as a goal-oriented project by which one aims to get rid of desires one cannot satisfy. However, he continues, such adaptation of preferences can be autonomous and indeed even more so than obsessively clinging on to desires that will necessarily remain frustrated. The point Sandven makes here appears plausible. Yet even his preferred account of what happens in mourning does not entail that the preference change involved in it is unconscious; the question would rather seem to be about intentionally and knowingly – though not necessarily in a very goal oriented fashion – working through the feelings of loss and frustration one faces. Accordingly, as he indeed makes clear, Sandven is saying rather that the distinction Elster makes between intentional, or goal-oriented, and unconscious adaptation of preferences is overly sharp than that unconscious preference adaptation is always compatible with autonomy.

12.4 Is the Requirement of Deliberation Overly Demanding?

Yet someone could maintain that the problems that adaptive preference change would seem to cause for autonomy are indeed merely apparent. If only choices based on deliberation (cf. the section *Adaptive Preference Change and Understanding* above) are autonomous, it seems that only a small part of the choices that competent adults normally make count as autonomous. Indeed, according to recent results of behavioral, cognitive, and neurosciences research and the newly emerged "paradigm of the adaptive unconscious," our everyday actions usually take place unconsciously, in an automatic manner (cf., e.g., Bargh and Chartrand 1999; Wilson 2002; and Hassin, Uleman and Bargh 2005). Accordingly, it could be suggested that requiring that only deliberate choices can be autonomous would be overly stringent. And hence, even though it is based on an adaptive preference, the conclusion here could be that the request for euthanasia presented by the patient in our main example should be deemed autonomous.

It indeed is plausible that in our busy everyday lives we often act rather unthinkingly and automatically. Yet as has been argued, for example, by Sie (2009), – and in a way reminiscent of how Hare conceived of "the levels of moral thinking"[13]– rather than being complete automatons, we do use our rational faculties to reflect on how we ought to act when we have the opportunity to do so.[14] Accordingly, requiring that *some* decisions can be autonomous only if they are based on deliberation would not be too demanding (see also, e.g., Beauchamp 2005). When the question is about choosing between, say, different flavors of ice cream, it would indeed normally be overly severe to require that the choice can be autonomous only if it is based on reflection of all of the pertinent matters. But when the choice is as important as that of deciding whether or not to continue living, the requirement that it ought to be based on deliberation does not appear overly demanding.

---

[13] Hare (1981) distinguished between intuitive and critical levels of moral thinking. According to him, at the critical level, when we have the time and other necessary resources for doing this, we reflect on moral issues and choose the moral principles which we then more or less intuitively apply in our everyday moral life. When the moral principles we have chosen at the critical level do not help us in the concrete situations facing us, we return to the critical level to solve the concrete moral problems troubling us, when we have the opportunity to do so.

[14] As some compatibilists about free will do, it can also be maintained that even if our belief that we are deciding was totally determined by physical factors beyond our awareness, we could still have reason to retain practices that apparently presuppose that we are not determined in that way. Accordingly, it is also arguable that retaining the conception of autonomy endorsed here might still be reasonable even if we were determined in that sense, which however need not be the one advocates of the "paradigm of adaptive unconscious" have in mind.

Other things being equal, in connection with a choice as significant as that, the patient himself must be as sure as possible that the decision he makes accords with his autonomous goals; otherwise he may end up dead though that is not what he autonomously desires. And securing that his request for euthanasia is congruent with his autonomous goals presupposes that the patient explicates to himself the reasons he has for it and, in light of them, assesses whether the choice really accords with those aims. As the patient's request is based on an unconsciously adapted preference, it does not meet this requirement. Furthermore, also the other parties involved – the physician who is responsible for treating the patient, the patient's close ones if such exist, etc. – need to be confident that the choice the patient makes really is autonomous. The only way in which they can acquire the required confidence is normally that of discussing the choice with the patient, determining whether he has reasons for making it in the way he does, and assessing whether the reasons and his choice accord with what in light of their best knowledge the patient autonomously wants.

12.5 Is Hypothetical Deliberation Sufficient for the Autonomy of the Euthanasia Request?

It could still be maintained that even if deliberation were necessary for the autonomy of a request for euthanasia, it need not be actual: hypothetical reflection is enough. Bruckner (2009, 317, 322, emphases in original), for example, describes an autonomous, or rational, adaptive preference as follows:

> …the notion of reflective endorsement in play here is not mere endorsement through a second-order preference, i.e., a preference for the preference one has. Rather, the endorsement is an all-in judgment that can conflict with a second-order preference. … An agent's adaptive preference is rational provided that *if* she *were* to examine the preference, *then* she *would* endorse it upon reflection.

As continuing the kind of life that he deems worthwhile is no longer possible for the patient, it could be suggested further, it is reasonable to believe that the patient would endorse his request for euthanasia were he to consciously reflect upon it. Accordingly, the conclusion here could be, even though it is based on an unconsciously changed preference the patient's request can be deemed autonomous after all.

However, the above presented reasons for the view that an autonomous request for euthanasia must be based on deliberation also entails that mere hypothetical reflection and endorsement is not sufficient to make the patient's request autonomous. As mentioned, that the patient can be sure that his request accords with his autonomous goals presupposes that he explicates his reasons for the request and, in light of them, assesses whether it accords with those goals. Moreover, as also already explained, that his physician and the possible other

parties concerned about him can be confident that his request for euthanasia accords with his autonomous aims presupposes that they consult the patient and discuss his reasons for the request with him. Accordingly, even if hypothetical reflective endorsement plausibly suffices for the autonomy of some preferences, when the question is about a choice as important as that of choosing between life and death, autonomy presupposes actual reflective endorsement.

Yet the kind of analysis that Bruckner advocates is still of importance in the context of the request under discussion. As it does not rule out autonomous acceptance of preferences that initially came about through adaptation, the conception of self-government adopted here allows that a patient can autonomously accept his adaptive preference. The patient is self-governing in endorsing the adaptive preference if the endorsement meets the criteria of autonomy explicated above. His preference would still be shaped by his options. But, when other things are being equal, the fact that he suffers from an incurable lethal illness would seem to be relevantly like the fact that he, say, is unable to survive without food and hydration. A person who does not come to terms with such realities will also be unable to make practical choices about his life autonomously. Accordingly, it appears that the problems that preference adaptation causes for autonomy could be overcome by critical reflection. And indeed it would seem that preference adaptation need not preclude the possibility of acquiring the kind of comprehension autonomy presupposes (see also Walker 1995, 465).

But then again, as already explained, sometimes the psychological drive to avoid frustration operative in preference adaptation can plausibly be powerful enough to undermine a patient's ability to come to terms with the realities of his life. When the psychological force is not that potent, then critical reflection can help in overcoming the problems that adaptive preference change causes the patient's autonomy.[15] But if the patient is unable to come to terms with his predicament, it is not possible for him to reflect on his adaptive preference. As already suggested, whether a given patient has the ability to engage in critical reflection of the pertinent kind is an empirical question that must be assessed on a case to case basis.

12.6 Should Autonomy Be Understood in More Rationalistic Terms?

Someone might grant that the above considerations show that preference adaptation threatens autonomy as it is conceived here, but still maintain that such

[15] An adaptive preference as it is here conceived is by definition a preference that is not based on actual critical reflection. Accordingly, if the endorsement is not merely hypothetical but actual, it would seem to be relevantly similar to what Elster calls character planning. Elster too allows that character planning and its resulting preferences can be autonomous (see Elster 1985, 127-128).

considerations are unsatisfactory precisely because they employ that particular understanding of autonomy. Even if we focused on the notion of autonomy that Beauchamp and Childress aim to explicate, the objection could go, the above interpretation of it in terms of intentional, informed, and voluntary action in accordance with one's own view of how one's life should proceed is not adequate. Besides the criteria to which Beauchamp and Childress refer, the criticism could proceed further, autonomy in that sense should also presuppose meeting some (additional) requirement(s) of rationality by which it is possible to deem, say, weak-willed, imprudent, and impulsive choices as heteronomous (cf., e.g., Walker 2008). And, as it remains unclear how preference adaptation would affect the autonomy of the patient's request for euthanasia when self-government is conceived of in such more rationalistic terms, the above argument is inadequate.

However, to the extent that they do not accord with a patient's own view of how her life should go, the conception of self-government adopted here entails that weak-willed, impulsive, etc. choices are not autonomous. Insofar as such choices are congruent with her autonomous conception of how her life should proceed – she has, say, autonomously come to accept her weakness for cigarettes – it is unclear why they should be deemed heteronomous. Consequently, contrary to what this possible objection implies, that we are able to deem such choices heteronomous when there is sufficient reason to do so would not seem to presuppose adding further requirements of rationality to the account of autonomy referred to here.

Moreover, at least as things are, the problem of what is and, perhaps to a lesser extent, also that of what is not rational remains quite controversial. That is reflected in, inter alia, the differing attitudes that people hold as regards a question such as whether it is rational to request euthanasia. Even if a particular answer to the question were rational according to some conception of rationality, a patient may autonomously reject the conception of rationality and, thus, also discard the answer that the conception of rationality entails to that question. And, as already suggested, enforcing, against her autonomous will, some rationally debatable answer to the question of whether or not to request euthanasia on a patient would be unjustifiable. Hence, for the time being at least, we should rather hold on to the liberal point of departure that patients who are self-governing in the above described sense ought to be allowed to decide by themselves whether or not to refuse life-saving treatment in light of what they consider to be the best thing to do. For these reasons, this possible criticism of the main argument of this article should not be accepted.

12.7 Conclusion

Is a request for euthanasia based on an adaptive preference autonomous? Insofar as the reflections presented above are plausible, the answer is no. Yet the above considerations also suggest that, when the psychological drive functioning in preference adaptation does not undermine a patient's ability to engage in critical

reflection, reflective endorsement can help with the problems that preference adaptation causes for autonomy. Unfortunately, the practical difficulties related to adaptive preference change do not seem easy. No established method for recognizing adaptive preferences is available[16] and a procedure that would allow us to demarcate adaptive preferences with absolute certainty is likely to remain an ideal. But, in terms of cases such as the one focused on here, since requests for euthanasia are usually presented when patients lack the option they would most want – continuing healthy lives – there is some reason to suspect preference adaptation whenever a patient requests euthanasia. If the request is based on a change in preference that is confusing even to the patient himself, the probability that the question is about preference adaptation is significant.

Of course, as they focus on patients' ability to make autonomous choices, current methods of competence assessment in medicine are plausibly already able to detect problems caused by preference adaptation. But, to better guard against the possibility that an adaptive preference is deemed autonomous, it would seem that healthcare providers should also inform their patients about the possibility that the patients' treatment choices are based on adaptive preferences and inquire whether they still, after an adequate cooling off period during which they have had sufficient time to reflect on the matter, are willing to endorse their choices. Ultimately, the question of what the most reliable way of identifying adaptive preferences is, and how the concrete problems that preference adaptation causes are best dealt with in medical practice are empirical rather than philosophical questions. This emphasizes the fact that, apart from medical ethicists, the question of how to deal with the phenomenon also merits attention from psychologists and social scientists working on empirical questions related to preference adaptation.[17]

References

Appelbaum, P.S., Lidz, C.W., Klitzman, R.: Voluntariness of Consent to Research: A Conceptual Model. Hastings Cent Rep. 39, 30–39 (2009)

Bargh, J.A., Chartrand, T.L.: The Unbearable Automaticity of Being. Am. Psychol. 54, 462–479 (1999)

Beauchamp, T.L.: Autonomy and Consent. In: Miller, F., Wertheimer, A. (eds.) The Ethics of Consent: Theory and Practice, pp. 55–78. Oxford University Press, New York (2010)

[16] It has been proposed that when a person unconsciously alters her preference because she has lost her (most) favored option she will also retrieve her original preference after regaining that option (cf., e.g., Elster 1985, 113). Yet adding the originally (most) favored option to a patient's option set is usually impossible in euthanasia cases and merely to say that it is, after all, open to the patient appears immoral.

[17] I thank Walter Glannon and Juha Räikkä for helpful comments on an earlier version of this chapter.

Beauchamp, T.L.: Who Deserves Autonomy and Whose Autonomy Deserves Respect? In: Taylor, J.S. (ed.) Personal Autonomy: New Essays in Personal Autonomy and its Role in Contemporary Moral Philosophy, pp. 310–329. Cambridge University Press, Cambridge (2005)
Beauchamp, T.L., Childress, J.F.: Principles of Biomedical Ethics. Oxford University Press, New York (2009)
Beauchamp, T.L., Davidson, A.I.: The Definition of Euthanasia. J. Med. Philos 4, 294–312 (1979)
Berges, S.: Why Women Hug their Chains: Wollstonecraft and Adaptive Preferences. Utilitas 23, 72–87 (2011)
Block, S.D.: Psychological Considerations, Growth, and Transcendence at the End of Life: The Art of the Possible. J. Am. Med. Assoc. 285, 2898–2905 (2001)
Borge, S.: The Myth of Self-Deception. Southern J. Philos 41, 1–28 (2003)
Bovens, L.: Sour Grapes and Character Planning. J. Philos 89, 57–78 (1992)
Bruckner, D.W.: Colburn on Covert Influences. Utilitas 23, 451–457 (2011)
Bruckner, D.W.: In Defense of Adaptive Preferences. Philos Stud. 142, 307–324 (2009)
Christman, J.P.: The Politics of Persons: Individual Autonomy and Socio-Historical Selves. Cambridge University Press, Cambridge (2009)
Colburn, B.: Autonomy and Adaptive Preferences. Utilitas 23, 52–71 (2011)
Colburn, B.: Forbidden Ways of Life. Philos Quart. 58, 618–629 (2007)
Courtwright, A.: Justice, Stigma, and The New Epidemiology of Health Disparities. Bioethics 23, 90–96 (2009)
Elster, J.: Sour Grapes: Studies in the Subversion of Rationality. Cambridge University Press, Cambridge (1985)
Ford, S., Lewis, S., Fallowfield, L.: Psychological Morbidity in Newly Referred Patients with Cancer. J. Psychosom Res. 39, 193–202 (1995)
Hare, R.M.: Moral Thinking: Its Levels, Method, and Point. Oxford University Press, New York (1981)
Harris, J.: Consent and End of Life Decisions. J. Med. Ethics 29, 10–15 (2003)
Hassin, R.R., Uleman, J.S., Bargh, J.A. (eds.): The New Unconscious. Oxford University Press, New York (2005)
Haworth, L.: Autonomy: An Essay in Philosophical Psychology and Ethics. Yale University Press, New Haven (1986)
Kraut, R.: What is Good and Why: The Ethics of Wellbeing. Harvard University Press, Cambridge (2007)
Levy, N.: Self-Deception and Moral Responsibility. Ratio 17, 294–311 (2004)
May, T.: The Concept of Autonomy in Bioethics: An Unwarranted Fall from Grace. In: Taylor, J.S. (ed.) Personal Autonomy: New Essays on Personal Autonomy and its Role in Contemporary Moral Philosophy, pp. 299–309. Cambridge University Press, New York (2005)
Mele, A.R.: Twisted Self-Deception. Philos Psychol. 12, 117–137 (2009)
Mill, J.S.: The Subjection of Women. Hackett Publishing Company, Indianapolis (1988)
Mill, J.S.: On Liberty. Penguin Books Ltd., Harmondsworth (1985)
Nussbaum, M.: Adaptive Preferences and Women's Options. Econ. Philos 17, 67–88 (2001)
Olsaretti, S.: The Concept of Voluntariness – A Reply. J. Polit Philos 6, 112–121 (2008)
Olsaretti, S.: Freedom, Force and Choice: Against the Rights-Based Definition of Voluntariness. J. Polit Philos 6, 53–78 (1998)

Sandven, T.: Autonomy, Adaptation, and Rationality: A Critical Discussion of Jon Elster's Concept of "sour Grapes," Part I. Philos Soc. Sci. 29, 3–31 (1999a)
Sandven, T.: Autonomy, Adaptation, and Rationality: A Critical Discussion of Jon Elster's Concept of "Sour Grapes," Part II. Philos Soc Sci 29, 173–205 (1999b)
Secker, B.: The Appearance of Kant's Deontology in Contemporary Kantianism: Concepts of Patient Autonomy in Bioethics. J. Med. Philos 24, 43–66 (1999)
Sen, A.: Gender Inequality and Theories of Justice. In: Nussbaum, M., Glover, J. (eds.) Women, Culture, and Development, pp. 259–273. Oxford University Press, New York (1995)
Sie, M.: Moral Agency, Conscious Control, and Deliberative Awareness. Inquiry 52, 516–531 (2009)
Walker, J.D.: Liberalism, Consent, and the Problem of Adaptive Preferences. Soc. Theory Pract. 21, 457–471 (1995)
Walker, R.L.: Medical Ethics Needs a New View of Autonomy. J. Med. Philos 33, 594–608 (2008)
Wilson, T.D.: Strangers to Ourselves: Discovering The Adaptive Unconscious. Belknap Press, Cambridge (2002)
Zimmerman, D.: Sour Grapes, Self-Abnegation and Character Building. Monist 86, 220–241 (2003)

Author Information

Bruckner, Donald W.
Assistant Professor of Philosophy, Pennsylvania State University, US. Representative publications include: Present Desire Satisfaction and Past Well-Being *Australasian Journal of Philosophy* (2013) 91:15-29, Against the Tedium of Immortality *International Journal of Philosophical Studies* (2012) 20:623-644, Colburn on Covert Influences *Utilitas* (2011) 23:451-457, and In Defense of Adaptive Preferences *Philosophical Studies* (2009) 142:307-324.

Colburn, Ben
Lecturer in Philosophy, University of Glasgow, UK. Representative Publications include: Autonomy and Adaptive Preferences *Utilitas* (2011) 23:52-71, Anti-Perfectionisms and Autonomy *Analysis* (2010) 70:247-256, *Autonomy and Liberalism* Routledge (2010), and Forbidden Ways of Life *Philosophical Quarterly* (2008) 58:618-629.

Dorsey, Dale
Associate Professor of Philosophy, University of Kansas, US. Representative Publications include: *The Basic Minimum: A Welfarist Approach* Cambridge University Press (2012), Subjectivism without Desire *Philosophical Review* (2012) 121: 407-442, Preferences, Welfare, and the Status-Quo Bias *Australasian Journal of Philosophy* (2010) 88:535-554, and Three Arguments for Perfectionism *Noûs* (2010) 44:59-79.

Grüne-Yanoff, Till
Associate Professor of Philosophy, Royal Institute of Technology, SWE. Representative publications include: Preference Change and Conservatism: Comparing the Bayesian and the AGM Model of Preference Revision *Synthese* (forthcoming), Old Wine in New Casks: Libertarian Paternalism Still Violates Liberal Principles *Social Choice and Welfare* (2012) 38:635-645, Models as Products of Interdisciplinary Exchange: Evidence from Evolutionary Game Theory *Studies in History and Philosophy of Science* (2011) 42:386-397, and *Preference Change: Approaches*

from Philosophy, Economics and Psychology Springer (2009) (Edited with Sven Ove Hansson).

Hanna, Jason
Assistant Professor of Philosophy, Northern Illinois University, US. Representative publications include: Consent and the Problem of Framing Effects *Ethical Theory and Moral Practice* (2011) 14:517-531, Paternalism and Impairment *Social Theory and Practice* (2011) 37:434-461, Revisiting Child-Based Objections to Commercial Surrogacy *Bioethics* (2010) 24:341-347, and Getting Lucky, Getting Even, or Getting Away with (Attempted) Murder: The Punishment of Failed Attempts *Public Affairs Quarterly* (2007) 21:109-123.

Hansson, Sven-Ove
Professor of Philosophy, Royal Institute of Technology, SWE. Representative publications include: The Harmful Influence of Decision Theory on Ethics *Ethical Theory and Moral Practice* (2010) 13:585–593, Objective or Subjective 'Ought' *Utilitas* (2010) 22:33-35, The False Dichotomy between Coherentism and Foundationalism *Journal of Philosophy* (2007) 104:290-300, and *The Structure of Values and Norms* Cambridge University Press (2001).

Khader, Serene J.
Assistant Professor of Philosophy, Stony Brook University, US. Representative publications include: Must Theorising about Adaptive Preferences Deny Women's Agency *Journal of Applied Philosophy* (2012) 29:302-317, *Adaptive Preferences and Women's Empowerment* Oxford University Press (2011), Adaptive Preferences and Procedural Autonomy *Journal of Human Development and Capabilities*(2009) 10:169-187, and When Equality Justifies Women's Subordination: Irigaray's Critique of Equality and the Fathers' Rights Movement *Hypatia*(2008) 23:48-74.

Rocha, James
Assistant Professor of Philosophy, Louisiana State University, US. Representative publications include: Pre-Emptive Anonymous Whistle-Blowing *Public Affairs Quarterly* (2012) 26:257-272 (with Edward Song), Autonomous Abortions: The Inhibiting of Women's Autonomy through Legal Ultrasound Requirements *Kennedy Institute of Ethics Journal* (2012) 22:35-58, Autonomy within Subservient Careers *Ethical Theory and Moral Practice* 14:3 (2011), 313-328, and The Sexual Harassment Coercive Offer *Journal of Applied Philosophy* (2011) 28:203-216.

Räikkä, Juha
Professor of Philosophy, University of Turku, FIN. Representative publications include: The Ethics of Alien Attitudes *The Monist* (2012) 95:511-532 (with Saul Smilansky), On Political Conspiracy Theories *Journal of Political Philosophy* (2009)

17:185-201, Global Justice and The Logic of the Burden of Proof *Metaphilosophy* (2005) 36:228-239, and On Irrational Guilt *Ethical Theory and Moral Practice*(2005) 7:473-485.

Schermer, Maartje
Professor of Philosophy of Medicine, ErasmusMC University Medical Center, NL. Representative publications include: Health, Happiness and Human Enhancement – Dealing with Unexpected Effects of Deep Brain Stimulation *Neuroethics* (forthcoming), Psychopharmaceutical Enhancers: Enhancing Identity? *Neuroethics* (2009) 2:103-111 (with Ineke Bolt), Enhancement, Easy Shortcuts, and the Richness of Human Activities *Bioethics* (2008) 22: 355-363, and *The Different Faces of Autonomy: Patient Autonomy in Ethical Theory and Hospital Practice* Kluwer Academic Publishers (2002).

Taylor, James Stacey
Associate Professor of Philosophy, The College of New Jersey, US. Representative publications include: *Practical Autonomy and Bioethics* Routledge (2009), Autonomy and Organ Sales, Revisited *Journal of Medicine and Philosophy* (2009) 34:632-648, Identification and Quasi-Desires *Philosophical Papers* (2005) 34:111-136, and Autonomy, Duress, and Coercion *Social Philosophy and Policy* (2003) 20:127-155.

Varelius, Jukka
Research Fellow, University of Turku, FIN. Representative publications include: Ending Life, Morality, and Meaning *Ethical Theory and Moral Practice* (forthcoming), Respect for Autonomy, Advance Directives, and Minimally Conscious State *Bioethics* (2011) 25:505-515, On the Prospects of Collective Informed Consent *Journal of AppliedPhilosophy* (2008) 25:35-44, and Autonomy, Well-Being, and the Case of the Refusing Patient. In: Pierce J, Randels G (eds.) *Contemporary Bioethics: A Reader with Cases* Oxford University Press (2009).