Münster Lectures in Philosophy 1

Simon Derpmann
David P. Schweikard *Editors*

Philip Pettit: Five Themes from his Work

Springer

Münster Lectures in Philosophy

Volume 1

Series editor
Department of Philosophy, Westfälische Wilhelms-Universität Münster,
Münster, Germany

More information about this series at http://www.springer.com/series/13531

Simon Derpmann • David P. Schweikard
Editors

Philip Pettit: Five Themes from his Work

Springer

Editors

Simon Derpmann
Department of Philosophy
University of Münster
Münster, Germany

David P. Schweikard
Department of Philosophy
University of Münster
Münster, Germany

Münster Lectures in Philosophy
ISBN 978-3-319-26101-0 ISBN 978-3-319-26103-4 (eBook)
DOI 10.1007/978-3-319-26103-4

Library of Congress Control Number: 2015958727

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

Over the last four decades, Philip Pettit has made a remarkable number of seminal contributions to a variety of fields of philosophy, ranging from metaphysics and the philosophy of mind and action to the philosophy of the social sciences, philosophy of law, ethics, and political philosophy. These works have not only advanced systematic thinking about the most pressing issues debated in the respective areas, they also – and this is particularly exceptional in a time of increasing compartmentalization of philosophy and specialization of scholars – display a continuing quest for systematic coherence. Moreover, especially in his recent work in republican political philosophy, Pettit has demonstrated a thorough concern for the historical depth of conceptual issues and arguments, thus situating his approach to the theories of freedom and democracy in a particular tradition he seeks to revive and update.

Pettit's work has received considerable attention from scholars around the world. This volume adds to the discussion, reception, and interpretation of his work on a range of issues from almost all areas of philosophy. It is based on the 16th *Münster Lectures in Philosophy*, hosted by the Department of Philosophy, Westfälische Wilhelms-Universität Münster, and held in October 2012. The volume comprises the evening lecture *Freedom and Other Robustly Demanding Goods*, the proceedings of a two-day colloquium during which groups of junior faculty and students presented critical comments on aspects of Pettit's work, and a synoptic reply essay by Philip Pettit.

First and foremost, we would like to thank Philip Pettit for coming to Münster, for delivering the *Münster Lecture* in 2012, and especially for the engaging discussions of his work. We are also very grateful for the support we received from the Department of Philosophy. We owe a special acknowledgment to our colleagues and students for the work they have put into their contributions to the colloquium, as well as to the many helping hands during the event. Finally, we thank Raphael

Hüntelmann for his sponsorship of this and many previous *Münster Lectures* and Lucy Fleet at Springer both for seeing to it that this valuable tradition will be continued and for supporting the finalization of this volume.

Münster, Germany
August 2015

Simon Derpmann
David P. Schweikard

Contents

Part I Lecture

Part II Colloquium

Part III Reply Essay

Abbreviations

CM	Pettit, Philip (1993) The Common Mind: *An Essay on Psychology, Society and Politics*. New York: Oxford University Press.
CP	Pettit, Philip (1997) The Consequentialist Perspective. In: *Three Methods of Ethics*: *A Debate*. ed. Baron, Marcia, Philip Pettit, and Michael Slote. Oxford: Blackwell.
DRA	Philip Pettit (2004) “Descriptivism, Rigidified and Anchored,” *Philosophical Studies* 118, pp. 323–38.
GA	List, Christian and Philip Pettit (2011) *Group Agency: The Possibility, Design and Status of Corporate Agents.* Oxford: Oxford University Press.
JD	Pettit, Philip (2007) Joining the Dots. In: *Common Minds: Themes from the Philosophy of Philip Pettit*, eds Brennan, Geoffrey, Robert Goodin, Frank Jackson and Michael Smith. Oxford: Oxford University Press: 215–344.
JF	Pettit, Philip (2014) *Just Freedom: A Moral Compass for a Complex World.* New York W.W. Norton and Co.
NJD	Braithwaite, John and Philip Pettit. 1990. *Not Just Deserts: A Republican Theory of Criminal Justice*. Oxford, Oxford University Press.
OPT	Pettit Philip (2012) *On the People’s Terms: A Republican Theory and Model of Democracy.* Cambridge, Cambridge University Press.
R	Pettit, Philip (1997) *Republicanism: A Theory of Freedom and Government.* Oxford, Oxford University Press.
TF	Pettit, Philip (2001) *A Theory of Freedom: From the Psychology to the Politics of Agency.* Cambridge and New York: Polity and Oxford University Press.

Part I
Lecture

Chapter 1
Freedom and Other Robustly Demanding Goods

Philip Pettit

1.1 Robustly Demanding Freedom

1.1.1 Freedom as Non-frustration

What exactly is required if you are going to enjoy freedom in any choice between certain options? This may be a choice, at the extreme of triviality, between walking to the left, walking to the right, and staying put. Or it may be a choice at the other end of significance between going to the right in politics, going to the left or holding to the center. Or of course it may be a choice between having tea or coffee for breakfast, going to the theater or a football game, studying philosophy or science at University. For ease of illustration, let us just assume that it is a choice of the kind that we would generally want our society to protect. Unlike a choice between doing another harm or not, it figures among the basic liberties that all can enjoy at once under a fair rule of law (OPT; JF).

One story as to what freedom in such a choice requires can be traced, like much else in the realm of political theories, to Thomas Hobbes, the seventeenth century English philosopher. In his monumental *Leviathan,* published in 1651, he offered us an account of what it is to enjoy freedom in certain choices or, in a curious use of

I learned a great deal in discussing this paper at the Muenster conference in October 2012, where it was first presented, and in other venues too: in particular, the Victoria University of Wellington and the Australian National University. I am particularly grateful for the helpful comments on an earlier draft received from Simon Derpmann and David Schweikard. The paper explores one among the many issues raised by the robustness of many moral values. For a fuller discussion, see Pettit (2015).

P. Pettit (✉)
University Center for Human Values, Princeton University, Princeton, NJ, USA

School of Philosophy, Australian National University, Canberra, Australia
e-mail: ppettit@princeton.edu

S. Derpmann, D.P. Schweikard (eds.), *Philip Pettit: Five Themes from his Work*,
Münster Lectures in Philosophy 1, DOI 10.1007/978-3-319-26103-4_1

the term, what it is to be a freeman.[1] He wrote: ‘a freeman is he that, in those things which by his strength and wit he is able to do, is not hindered to do what he has a will to’ (Hobbes 1994, Ch 21). We may ignore the middle clause in the sentence, which tells us that free choices are restricted to choices within the capacity—the strength and wit—of the agent. The point to focus on is his suggestion that it is enough to be free in a choice that you manage to do—you are not blocked from doing—that which you have ‘a will to’ do: that which you prefer among the options before you.

This suggestion means that to enjoy freedom between options in choices like those illustrated earlier—schematically, freedom in any choice between options X, Y and Z—all that has to happen is that among these options you can get to enact the one that you prefer. Freedom in the choice is indistinguishable, on this account, from preference-satisfaction: that is, satisfaction of the preference you actually hold over the alternatives at issue.

Hobbes did not endorse this conception of freedom just by way of an unremarked implication of his definition of a freeman. He was taken to task about his claim in a famous exchange with a contemporary bishop and philosopher. Bishop Bramhall suggests in that exchange that if you are considering whether or not to play tennis—we assume a willing partner—and you decide against doing so, then you may still have been wrong to think that you had a free choice. After all, unbeknownst to you, someone may have shut the door of the (indoor or ‘real’) tennis court against you. Hobbes is undaunted by the argument, asserting that for anyone in your position ‘it is no impediment to him that the door is shut till he have a will to play’ (Hobbes and Bramhall 1999, 91).

1.1.2 *Freedom as Robust Non-frustration: Berlin*

In this analysis of freedom, as on so many issues in political theory, Hobbes’s influence was enormous. It continues today in the prevailing economic habit of identifying freedom with preference-satisfaction. But it is close to demonstrable that this analysis of free choice does not fit with our deeply ingrained habits of thinking: that while it may represent a possible conception of free choice, it is not the conception that most of us actually endorse in our thinking about such matters. To identify freedom with preference-satisfaction is to embrace an absurdity, as Isaiah Berlin (1969, xxxix) has pointed out (Pettit 2011; OPT).[2]

[1] In using this term, Hobbes was almost certainly wanting to deny it to the republican tradition that he opposed. As we shall see later that tradition emphasized the importance of being a liber who is sui juris—being, as this idea was translated, a freeman—in order to enjoy freedom.

[2] Berlin most clearly focuses on this point in the 1969 introduction to the collection in which his 1958 lecture on ‘Two Concepts of Liberty’ was published and acknowledges doing so as a result of criticism by an anonymous reviewer of the 1958 lecture in the Times Literary Supplement. That reviewer, it appears, was Richard Wollheim. I am grateful to Albert Weale and Jonathan Wolff for throwing light on this for me.

Suppose that you are in prison and wish to live in the outside world. That means that by Hobbes's account, you are unfree in the choice between living behind bars and living outside. But according to that account you will be able to make yourself free in this choice if you can only come to prefer the option that is available to you: continuing to live in prison, say for another three years. And of course it may be within your power to get your preferences to shift in that direction. Suppose you reflect routinely on the good points about prison in comparison to life outside: a roof over your head, regular meals, the chance to read lots of books, and so on. There is a good chance in that event that you will cause your preferences to adapt and that you will become more or less happily resigned to prison. And if you succeed in this then, according to the Hobbesian view, you will have made yourself free.

By Berlin's reckoning, and surely he is not on his own, this is quite absurd. In order to win freedom in a choice, it appears that you have to shape the world so that it allows you to choose as you wish. It cannot be enough that without changing anything in the world, you manage to change your wishes. As Berlin (1969, 139) expresses the absurdity: 'I need only contract or extinguish my wishes and I am made free'. He puts the point even more forcefully in a later comment. 'To teach a man that, if he cannot get what he wants, he must learn to want only what he can get may contribute to his happiness or his security; but it will not increase his civil or political freedom' (Berlin 1969, xxxix).

Think of each option in a choice as a door, where an option is available just when the door is open (Berlin 1969, xlviii). What Berlin's argument from absurdity shows is that it is not enough for freedom of choice in the ordinary sense that the door you actually choose to push on happens to be open. It must also be the case that for any other option in the choice, for any door that you might have pushed on instead, it too is open. Suppose we are considering whether you are free in a choice between options, X, Y and Z. If you are truly to count has having a free choice in the actual world, then you must not be hindered in the actual world where you prefer X but equally you must not be hindered in the possible worlds where you prefer Y or Z. You must get what you actually want but it must also be the case that you would have gotten one of the other options had you wanted it instead.[3] All doors must be open.

Although he does not put it this way, the lesson from Berlin is that freedom in a choice is a robustly demanding good. It requires that in the actual world where you prefer X, you get X: you enjoy non-frustration by another. But it also requires that in the possible worlds where you prefer Y or prefer Z, you get Y or Z: you enjoy non-frustration in those possible worlds, as well as enjoying it in the actual world.

[3] Strictly, there is a problem in saying that to be free in the choice of X, it must be the case that you could have chosen the alternative, Y, had you wanted to—had you preferred that option. This condition might be incapable of fulfillment because you are the sort of person who would only want to do Y if it was not an available option; the possibility will be salient from Groucho Marx's quip that he would only want to join a club that would not accept him as a member. The problem can be overcome if what is required is that you could have chosen Y had you tried to do so, where it is not required in that eventuality that you actually prefer Y. For expressive convenience, I shall ignore this complication in the text. I am grateful to Lara Buchak for alerting me to the problem.

You actually enjoy non-frustration and you would have continued to enjoy non-frustration, no matter which among the options you happened to prefer.

Freedom on this account is robustly demanding, because it requires the presence of a less demanding good, non-frustration, not just in the actual world, but in various possible worlds in which your preferences over the options are different. The worlds over which non-frustration must continue in order for you to enjoy actual freedom at our hands do not extend to all possible worlds where you form a different preference between the options, only to worlds that resemble the actual world in a distinctive manner. They are sufficiently like the actual world for the considerations that make freedom valuable to remain relevant to how we ought to treat you and to retain a greater weight than competing concerns. And they are sufficiently alike for our natures not to be transformed or corrupted there; in that sense they are modest variations on the actual world. You could hardly complain that you do not enjoy freedom at our hands, because we would interfere with you if there were good reasons to do so or if we had turned into monsters.

But however restricted in these ways, the worlds over which we must grant you non-interference in Berlin's sense, if you are to enjoy freedom at our hands, are not limited to relatively probable worlds. It might be very unlikely indeed that you should ever prefer to have Z rather than X or Y but if you are to be free in the X-Y-Z choice, then you must still enjoy non-frustration in that Z-preferring world. It must still be the case that if you had preferred Z, however improbable that is, you would not have been subject to frustration in realizing Z.

1.1.3 Freedom as Robust Non-frustration: Republicanism

Subject to the restrictions of relevance, weight and modesty, then, you enjoy freedom in a certain choice just to the extent that you can choose as you wish, regardless of what you wish. You can choose as you wish, avoiding frustration, regardless of whether you prefer X or prefer Y or prefer Z. This already shows that freedom is a robustly demanding good, in the sense defined, but it is worth remarking that on the republican tradition of thinking, as I and others have argued elsewhere, freedom is even more robustly demanding than this suggests. Subject to the three restrictions, your actual freedom in a choice requires not just that you should be able to choose as you wish, regardless of how you wish to choose, but also that you should be able to choose as you wish, regardless of how others wish that you choose (Pettit R; OPT; JF). Not only must all the doors in the choice be open; there must be no doorkeepers who have the power to close them at will, should they take against you.

According to this strengthened view, the non-frustration required for your actual freedom has to remain in place, not just across worlds that vary in what you prefer to choose, but also in worlds that vary in what others prefer that you choose. There are a number of arguments for assigning this richer robustness to the value of freedom but let it suffice here to mention just one. This is that it is natural to think that you are your own boss in a certain type of choice—and that you enjoy freedom of choice in that sense—only insofar as you are not subject to the will of another as to how you should choose. The idea is that you should not have to depend on the will of another being favorable to your choosing as you wish in order to be able to choose as you wish; you should not be able to choose as you wish only because some other gives you permission to do so. Did you depend on getting the leave or permission of another in order to make a choice, then that person's will would be in ultimate charge of what you do, not your own. And this would be true, even if it was extremely unlikely that the other would deny you permission. No matter how much goodwill they bore towards you, that person would be in control and you would not be properly free.

This thought is to be found in Roman republicans like Cicero and Livy, in modern thinkers like Machiavelli, Harrington and Locke, and in a batch of enlightenment thinkers, including Kant. If we endorse it, as I think there is good reason to do, then we must think that freedom requires non-frustration in a maximally robust sense. It requires that you should escape the frustration of another in actually exercising your choice and it requires that you would escape that frustration, regardless of what you wanted to do and regardless of what any other wanted you to do. In exercising the choice, you must enjoy the status, as it was described in Roman law, of being a *liber* or free person who is *sui juris*: under your own jurisdiction (Skinner 1998). You must not be subject to the power of any other in that respect; you must enjoy freedom in a sense in which it requires non-domination: not being exposed to the *dominatio* of any *dominus* or lord in your life.[4]

Whether or not we go along with this republican radicalization of Berlin's idea, we must admit that freedom of choice is still a robustly demanding ideal. It requires the presence of the less demanding good of non-frustration in the actual world but also in a range of possible worlds. By all accounts the only relevant possible worlds are those modest variations on the actual world in which the reasons for why you should be able to choose as you prefer in the choice on hand retain their relevance and weight. According to Berlin's view, the only such worlds crucial for whether you are actually free are those where you change your mind about what you want to choose. According to the republican view they extend also to worlds where others change their minds about how they want you to choose.[5] Under each view, freedom is a value or ideal that makes robust, not just actual, demands. Only the Hobbesian alternative would hold otherwise.

[4] Conceptualizing freedom in this way, the republican tradition argues that under the law each citizen should be protected against domination in a range of choice that each can exercise and enjoy at the same time that others exercise and enjoy it. For a classic statement of that thought, see Kant's (1996) Metaphysics of Morals and for more recent efforts see (OPT, JF).

[5] I ignore one extra dimension of robustness. This is that either view of freedom would require you to enjoy non-interference with what you do, no matter how in particular you decide to do it: no matter which foot you put forward, for example, in deciding to turn left rather than right.

1.2 Other Robustly Demanding Goods

1.2.1 Robustly Demanding Attachments

Once we see the modal structure that freedom displays, it should be clear that there are lots of values or goods that we treasure in human life that embody the same sort of structure. I start with the good of love—or by extension, friendship or collegiality or solidarity—that I may confer on you and you, reciprocally, on me. There are many different styles of love, of course, as there are many different styles of friendship and other relationships, but I put aside such complications here in order to underline the point about common structure.

A good way of showing that love has the structure of a robustly demanding good is to begin with a play—perhaps one of the great comedies of all time—by Oscar Wilde. In *The Importance of being Earnest,* Jack Worthing uses the pseudonym 'Ernest' on his visits to London, as he wishes to retain a certain anonymity in the big city. Under that pseudonym he attracts the attentions of Gwendolen, the cousin of his friend, Algernon, and they fall in love.

Or do they? Gwendolen's attachment may not earn the name of love, since it transpires that it is only name-deep. As she explains in response to his confession of attachment: 'my ideal has always been to love some one of the name of Ernest. There is something in that name that inspires absolute confidence'—presumably, we are meant to suppose, the fact that it sounds like 'earnest'. And as if that were not sufficiently bewildering, she adds: 'The moment Algernon first mentioned to me that he had a friend called Ernest, I knew I was destined to love you'. Jack remonstrates with her, of course, explaining that he would much rather be called 'Jack'. But Gwendolen will have none of it, expanding with ever greater enthusiasm on the charms of 'Ernest'. 'It suits you perfectly. It is a divine name. It has a music of its own. It produces vibrations.'

Does Gwendolen really love Jack? Well, if she does, she has a strange way of thinking about it and that is part of what is so funny about Jack's predicament. What makes it even funnier is that he immediately wonders if he should be re-christened 'Ernest', as if that would put the situation right. The theme is amusing, because the passionate degree of love that Gwendolen declares for Jack fits ill with its turning on the fortuity of his name. We expect that if she loves him, then her attachment ought not to be contingent on the fact that, as she thinks, he is called Ernest. It ought to be more robust than that.

Wilde's comedy teaches us that while the good of love—the good that consists in enjoying the love of another—certainly requires the affectionate concern that Gwendolen declares for Jack, it also requires something more. If I love you then, as things actually are, I have to offer you due care. I have to register and respond to the stimulus of your needs and wishes in the partisan manner we expect of a lover. But the due care I offer you in this way must not be premised just on how you happen to be: it must be able to survive a variety of possible changes in you, among them the change or apparent change in the name you bear. If I love you, then not only must I offer care that would survive sickness as well as health, poverty as well as affluence, to cite the standard vows. I must also offer care—under the impact of suitable triggers—independently of how you currently look, what you currently do, or how you are currently called. Shakespeare already made the point in one of his sonnet (No 16). ‘Love is not love, Which alters when it alteration finds’.

We saw in the last section that the richer good of freedom pairs off with the thinner good of non-frustration and that what makes freedom richer is that it requires non-frustration, not just in the actual world, but also in a range of associated possible worlds. These worlds allow changes in your preference and mine over what you choose and they resemble the actual world sufficiently to satisfy constraints of relevance, weight and modesty. They are modest variations on the actual world in where there are still relevant, suitably weighty reasons for me to grant you non-interference. We are in a position now to see that as freedom relates to non-frustration in this way, so love relates to the due care—the exercise of care under suitable prompts—that lovers typically provide for those they love.

Love is a rich good that requires the robust, not just the actual presence of the thin good of such care. Consider the range of possible worlds that are sufficiently close to the actual one to satisfy constraints of relevance, weight and modesty. There are still reasons of love for showing you care in those scenarios: you have not turned out to be a covert murderer, for example. Those reasons still retain their weight in relation to other considerations: showing you care, for example, will not cause the death of a third party. And in those worlds I have not become utterly alienated from you, as can happen in the best of relationships. If you are actually to enjoy my love, it must be the case that you would enjoy my due care under those sorts of possible scenarios as well as enjoying it just as things happen to be. And this is so, even if some of those scenarios are pretty improbable.

Thus the fact that you were less charming, less healthy or just older ought not to change how I would treat you. And nor should the fact that I might find it less convenient or congenial to offer you such treatment. Shakespeare may have gone over the top when he said that love constitutes ‘an ever-fixed mark, That looks on tempests and is never shaken’. But the thought is still on the right track. The actual love I bear you requires me to be ready to offer you suitable care and concern over a range of possible scenarios, some of them quite improbable, where you lose some of your charms and things make it more difficult or costly for me to give you loving care.

As it is with love, so it is with a range of other attachments, in particular with the attachment of friendship. If I claim to be your friend but would only deliver the concern that friendship requires so long as that answered to my advantage or conformed to my inclination, then I would not deserve the name of friend. I would be, as we say, merely a fair-weather friend: not the real thing but only a simulacrum. Like love, friendship and other attachments require the delivery of a suitable thinner good and, in particular, its delivery across a range of merely possible scenarios, not just in the actual world. They are rich goods in the sense that they impose demands that bear on how I actually deal with you but also on how I would deal with you under a swathe of variations on the actual world.

1.2.2 Robustly Demanding Virtues

The robustly demanding structure of attachments like these is also replicated in values of a less particularistic kind: values or goods that we enjoy at the hands of many different individuals, not just those to whom we relate as lovers or friends or colleagues or whatever. Consider the honesty that you and others may enjoy in your dealings with me, particularly when you seek information. Or consider the justice you and others may enjoy at my hands insofar as I satisfy any claims that you have against me. Or consider the trustworthiness that you and others may enjoy when you rely on me, as a matter of shared awareness, not to let you down: not to renege on a promise, for example, or not to leave you otherwise in the lurch. As attachments require me to provide robustly for your enjoying a thin good of some corresponding kind, so the same is true of what virtues of this less personal kind require.

In order for you to enjoy my honesty, I must certainly tell you the truth as things actually are. You would not enjoy my honesty, however, unless I was disposed to tell the truth not just under those circumstances but also under variations where it was less convenient, for example, to tell you the truth. Consider those modest variations on the actual world—variations in which I am not utterly corrupted—that are sufficiently similar to actual circumstances for reasons of honesty to remain relevant and to continue to outweigh rival considerations. If you are to enjoy my honesty then it had better be the case that I would tell the truth, not just in the actual world, but under all such variations. Only in that case can I count as an honest person and only in that case can you enjoy the good that my honesty provides: the good of my honesty, as we say.

What holds for honesty holds also for justice. In order to relate to you in justice I must be prepared to honor your claims against me, whatever these are taken to be. But I must be prepared to honor them robustly, across a variety of possible circumstances, including the actual world, not just when things happen contingently to be propitious: not only, for example, when it happens that satisfying your claims is independently in my interest or corresponds with my independent inclinations. To use our emerging formula, if I am to give you justice then I must be prepared to act on reasons of justice—reasons related to your claims against me—not just in the actual world but in modest variations on the actual world that are sufficiently similar for reasons of justice to remain relevant and weighty.

Thomas Hobbes (1994, Ch 15) underlines the point in the case of justice when he says that a just man is ‘he that taketh all the care he can that his actions may all be just, an unjust man is he that neglecteth it’. He means that the just man is ready to satisfy the claims of others, however things turn out to be: even, for example, should they ensure that satisfying the claims of others will be onerous or costly. The point he makes is of ancient pedigree. It appears already in the *Digest of Roman Law,* produced under the Emperor Justinian in the sixth century BCE*: Justitia est voluntas constans et perpetua jus suum cuique tribuendi;* ‘Justice is the steady and enduring will to render unto everyone his right’ (Watson 1985, I.1.10). The important point in this principle is that justice does not consist just in giving everyone his right or his due, as the translation has it, but in giving everyone his due out of an unswerving will: that is, in a robust or reliable manner.

The points illustrated with honesty and justice apply to a range of impersonally rich goods that we bestow on one another. The final example I mention is the trustworthiness I may display in proving reliable in any case where it is manifest to me, perhaps as a result of an explicit promise or a mutual understanding, that you are relying on me in some salient fashion not to let you down. It will not be enough for me to give you my trustworthiness—for me to earn your trust—that I actually prove reliable, or am likely to prove reliable, because that answers to other interests. I must be ready to prove reliable—ready to deliver that thin good—across an open range of scenarios. More specifically, I must be prepared to do so, not just in the actual world, but in any modest variations scenarios that are sufficiently similar for reasons of trust to remain relevant and weighty.

1.2.3 Beyond Moral Behaviorism

These observations ought to be enough to support the claim that many other values or goods in human life display the same robustly demanding structure that freedom exemplifies. But before taking that point to be established, and before moving to the discussion of its relevance for consequentialism, it may be useful if I address one important objection. Opponents of the line I take may say that in being a lover or a friend, or in being honest or just or trustworthy—for short, in espousing and exercising such dispositions or attitudes—I do not do anything for you and it is misleading to say that I provide you with the relevant good. They may argue that it is only in the provision of the thin benefits that correspond to those values—only in providing due care or concern, only in delivering truth-telling or claim-satisfaction or reliable behavior—that I actually do anything. In other words, they may defend a sort of moral behaviorism, according to which it is one thing to do good, it is quite another to have good dispositions or to exercise good dispositions. And with such a moral behaviorism in the background, they may say that what I have been discussing are conditions for an agent to be a good person, not conditions for an agent to provide certain goods for another.

Although the phrase has no currency, moral behaviorism is characteristic of a great deal of contemporary writing in the area of ethics or moral philosophy (Scanlon 2008). The position is well represented by Jonathan Bennett (1995, 49), who argues that ‘the basic concern’ of morality is with the behavior of agents, not their attitudes. Following Alan Donagan (1977), he makes a distinction between a first-order morality focused on behavior and a second-order morality focused on attitude. ‘First-order morality issues judgments of the type “It would be right for me to φ”, “He acted wrongly in φing”, and so on; second-order morality judges whether the person deserves credit or discredit–perhaps including praise or blame–for φing’ (46). For Bennett the attitudes out of which an agent acts ‘are relevant to judgments in the associated second-order morality but not to first-order judgments of wrongness’ (49). They need to be taken into account in determining how good the agent is but not in determining whether or not the agent’s action was a case of acting well or rightly.

Moral behaviorists would say that to speak, as we have been doing, of the goods I provide for you in possessing and manifesting the disposition of a lover or friend, or of someone honest or just or trustworthy is to confuse first-order with second-order morality. But this runs counter to the common sense that I put something good and valuable in your hands when I act as a lover or friend, an honest or just or trustworthy person: something in particular that I would not put in your hands if I displayed the same behavior but out of a different source. By their lights the helpful action of the friend is just the same action as the helpful action of an opportunistic acquaintance and calls to be judged by the same criteria of right and wrong. And this violates the common intuition that they are quite different acts, the one being performed out of genuine friendship and the other out of rank self-interest (Herman 2011).

Moral behaviorists presumably take this view to be non-optional, because of holding that action always has to be characterized just by its actual consequences, independently of the motive or disposition out of which it is performed. But action does not have to be characterized in that way. When I act as a friend and give you the benefit of my concern, then there is something I can be said to do that an opportunistic acquaintance could not be described as doing. I program or control for the production of that benefit, being of such a mind that no matter how circumstances turn out—within now familiar constraints—I produce the benefit in those circumstances. There is a world of difference between producing a benefit for you, no matter out of what motive or disposition, and programming for that benefit. And programming has as good a call as production to be treated as a form of action.

Programming for an effect, like producing an effect, is not confined to human agents. As Frank Jackson and I have argued elsewhere, the higher-order causes that we take to be operative in the natural, psychological or social world all have the feature that no matter how they are realized at lower levels—no matter, ultimately, what their microphysical character amounts to—they are generally succeeded by the effects of which we take them to be causes (Jackson et al. 2004, Pt 1). If we take the microphysically characterized factors to produce the relevant effects, we can say that the factors under their higher-level characterization program for those effects. They set things up so that the effects are more or less bound to follow. In this sense a rise in unemployment may program for a rise in crime, for example, as some criminologists have claimed. No matter how the rise in unemployment is realized, whether in the dismissal of this or that group or in this or that industry, the change is liable to trigger a corresponding rise in criminal activity.

One of our distinctive characteristics as agents is that we program in broadly this sense for most of the effects that we produce in the world and in the people around us. We are attitudinally such that no matter what perceptions and beliefs we form—no matter in that sense, what precise form our attitudinal configuration takes—we will adopt the steps required according to our attitudes for bringing about this or that effect. The very idea of my doing something intentionally—say, producing an effect, E—would make little sense except insofar I am disposed to act so that no matter which of a number of possible scenarios is realized—and so no matter which configuration of beliefs I instantiate—I bring about E. Do I seek a drink intentionally when I go to the fridge and remove the beer? Of course I do, for I program for getting a drink in the sense that within certain limits I am disposed, depending on my perception of circumstances, to take this or that initiative in order to secure a drink. Let the beer be on the counter, as I think, not in the fridge, and I will go to the counter. Or let it be the case, as I think, that there is only water around, not beer, and I will go for water.[6]

If we think that human action can be conceptualized so that it involves programming for events, as it surely can, then the divide that moral behaviorists see between the assessment of actions and the assessment of attitudes and agents takes on a different cast. This conceptualization allows us to say, in line with the merest common sense, that in acting as a lover or friend I program for producing due care and concern for your welfare; in acting as someone honest I program for telling the truth; as someone just I program for satisfying claims against me; and in acting as someone trustworthy I program for proving reliable in face of the expectations of others. And the actions that I am thereby ascribed deserve to be distinguished from—and of course rated more highly than—the actions characterized just by the thin benefits produced. I may be assessed as an agent for how far I possess the dispositions in question and, given that my manifestation of those dispositions may be thwarted, this assessment may come apart in some measure from the assessment of my actions. But this distinction puts much less on the side of agent-assessment, and much more on the side of action-assessment, than moral behaviorists would allow.

The upshot is clear. The moral behaviorist line is at best a non-compulsory way of keeping the books and, at worst, a highly counter-intuitive story to offer. There is no reason why we should allow it to undermine the observations that we have made in this section. We can continue to say, with a good conscience, that apart from the thin benefits I produce in any acts of friendship or honesty or whatever—the thin benefits that I might equally have produced out of self-interest—there are a variety of rich goods that I bring about in the manifestation of such dispositions.

1.3 The Implications for Consequentialism

According to consequentialism the right option to take in any choice, roughly, is that which promises to have the best results overall: that which promises to maximize the good. Let us assume, in a pluralistic version of consequentialism, that there

[6] There are two different ways in which I might program for an effect: say, to take a toy case, I might program for effect E, being disposed to take a suitable means: means M1 in circumstance C1, means M2 in circumstance C2…and means Mn in circumstance Cn. Suppose C1 transpires and I take M1 and achieve E. What might make it the case that I programmed for E? One is that I was disposed to choose M1 in C1, M2 in C2…and Mn in Cn, where this is explained by the fact that each was a way to achieve E. In this case, I programmed for achieving E by planning to achieve E: by forming a conditional plan for each C to adopt a corresponding M (Bratman 1987). The other way I might have programmed for E is by having been disposed at a higher level so that should C1 transpire, I would become disposed to take M1, should C2 transpire, I would become disposed to take E2…and should Cn transpire, I would become disposed to take M3; and all of this because of the connection with achieving E. I programmed in this case for achieving E without anything like planning being involved. I programmed for E in virtue of the structure of my dispositions, not in virtue of the content of any single disposition.

are many goods, however they are to be weighed against each other, and that among the goods that consequentialism should take into account are goods like those we have been considering here: goods like the freedom you can enjoy in certain choices, the love and friendship you can win at the hands of those near and dear, the honesty and justice and trustworthiness that you can savor in those with whom you routinely interact. The question I want to address in this final section is whether our discussion of such goods has implications for how we should think about the nature and the likely direction of consequentialist thought.

I believe that it has important implications for this topic, though I shall concentrate only on one. This implication bears on the role of attachment and virtue—and, by extension, law—in ethical theory. The view adopted here means that attachment, virtue and law are more intimately connected with doing good by consequentialist lights than is recognized in the literature.

Attachment is relevant to doing good, by the argument of this paper, insofar as it is only by my being attached to you as a lover or friend—only by my exercising a lover's or a friend's disposition—that I do the good that consists in giving you love or friendship. Virtue is relevant insofar as it is only by being honest or just or trustworthy with you—only by exercising those dispositions—that I give you the goods that go by the same names: my honesty or justice or trustworthiness. And law is relevant insofar as it is only by grace of our collectively setting up an appropriate law that we can individually give you the good of freedom in at least the republican sense. Under plausible assumptions, it is only if we are each bound by a protective law that you can choose as you wish within the space created by that law, regardless not only of what you wish to do but regardless also of what any one of us may wish that you should do.

There are two standard accounts of why attachment, virtue and law should be taken seriously from within a consequentialist frame—in particular a frame that allows for the plurality of goods—as contributing to the good. One of these is practical in character, the other epistemic, but it turns out that they do not exhaust the possibilities.

The practical approach is that suitable forms of attachment or virtue or law can each make it more likely than otherwise that we as individuals produce actual-world benefits like those we considered in each case: due care or concern, truth-telling or claim-satisfaction, or the non-frustration of choice. Attachments and virtues serve as dispositions whose presence triggers us to produce those benefits more frequently than if we lacked them. And laws serve as coercive forms of regulation that hold out threats which make it more likely that we will avoid frustrating the relevant choices of others.

The epistemic approach focuses on two aspects of attachment, virtue and law. First, the way in which they can sharpen our perceptions and make it more likely that we will identify opportunities where there are benefits to be produced. And second, the way in which they can impress themselves on others, providing an assurance that we will provide the benefits promised. My attachment as a friend, it will be said, is going to cue me to the triggers for providing various benefits for you, my friend, and is likely when recognized to reassure you about my behavior.

My virtue of honesty or justice is going to let me see more clearly the grounds in this or that case for telling the truth or satisfying various claims (McDowell 1979), and is likely when observed to provide others with a reassurance about my intentions. And the laws that protect this or that domain of choice—say, by identifying liberty of speech or association or religion—are bound to have analogous effects. They will provide me and others with a signal that that is an area where we should avoid frustrating the choices of others; and they will furnish each of us with a degree of assurance that almost everyone is going to comply.

The lesson of our discussion in earlier sections is that attachment and virtue and law are not only of practical and epistemic relevance in a consequentialist theory of the right. They are relevant also in a far more important, ontic or ontological respect. There are rich goods generated by our acting out of attachment or virtue or law and these contributions to the good are quite distinct from the contributions charted in the practical and epistemic stories.

Why describe the connection between attachment, virtue and law on the one side and the promotion of the rich good on the other as ontic or ontological? In brief, because it is impossible for me to act towards you out of the relevant disposition without conferring on you the corresponding rich good. To act out of the disposition of friendship towards you is just to give you my friendship, actually and robustly conferring due concern. To act out of the disposition of honesty, actually and robustly telling you the truth, is just to give you my honesty. And to individually respect a collectively imposed law—a constitutional provision, a statute or perhaps just a custom—that prohibits each of us from frustrating you, no matter how far we might wish to do so, is just to give you freedom.

If we adopt the practical or epistemic approach to attachment, virtue and law then as consequentialists we will think that they certainly have a high degree of causal-instrumental importance. But if we adopt the ontic approach that this paper would support, then we must give them a different sort of importance. They are not just means whereby important goods are contingently generated; they are means whereby those goods are necessarily made available. The connection between the dispositions and institutions, on the one side, and the goods they underwrite on the other is not a causal relationship or anything of that contingent kind. It is a constitutive relationship of the sort that exists between the antibodies in my blood and the immunity that they confer on me against certain diseases. Just as my antibodies make it the case that I am immune, so the dispositions out of which I act, and the laws under which I and others act, make it the case that you enjoy the robustly demanding goods associated with them.

And so to our conclusion. If as consequentialists we put a high premium on the rich goods charted here, then we are bound to take a very distinctive view of the importance of attachment, virtue and law. We can argue for putting these in place in a manner that is not contingent on their happening to serve the production of distinct thin benefits, in the fashion of the practical or epistemic approach. There will be a case for establishing them that derives from the constitutive manner in which they ensure the realization of the robustly demanding goods we have been charting here. This observation opens up many issues, for it calls for exploring aspects of consequentialism that connect it with virtue theory and other approaches long deemed inimical. The observation offers an indeterminate note on which to end but not perhaps a bad one. Better to point forward than to look back.

References

Bennett, J. 1995. *The act itself*. Oxford: Oxford University Press.

Berlin, I. 1969. *Four essays on liberty*. Oxford: Oxford University Press.

Bratman, M. 1987. *Intention, plans, and practical reason*. Cambridge: Harvard University Press.

Donagan, A. 1977. *The theory of morality*. Chicago: University of Chicago Press.

Herman, B. 2011. A mismatch of methods. In *On what matters*, vol. 2, ed. Derek Parfit and S. Scheffler, 83–115. Oxford: Oxford University Press.

Hobbes, T. 1994. *Leviathan*, ed. E. Curley. Indianapolis: Hackett.

Hobbes, T., and J. Bramhall. 1999. *Hobbes and Bramhall on freedom and necessity*, ed. Vere Chappell. Cambridge: Cambridge University Press.

Jackson, F., P. Pettit, and M. Smith. 2004. *Mind, morality, and explanation: Selected collaborations*. Oxford: Oxford University Press.

Kant, I. 1996. *Practical Philosophy*. Trans. M.J. Gregor. Cambridge: Cambridge University Press.

McDowell, J. 1979. Virtue and reason. *The Monist* 62: 331–350.

Pettit, P. 2011. The instability of freedom as non-interference: The case of Isaiah Berlin. *Ethics* 121: 693–716.

Pettit, P. 2015. *The robust demands of the good: An ethics of attachment, virtue and respect*. Oxford: Oxford University Press.

Scanlon, T.M. 2008. *Moral dimensions: Permissibility, meaning, blame*. Cambridge: Harvard University Press.

Skinner, Q. 1998. *Liberty before liberalism*. Cambridge: Cambridge University Press.

Watson, A. 1985. *The digest of justinian, four volumes*. Philadelphia: University of Pennsylvania Press.

Part II
Colloquium

Chapter 2
Rule-Following and A Priori Biconditionals – A Sea of Tears?

Amrei Bahr and Markus Seidel

2.1 Pettit's Solution to the Problem of Rule-following: Response-Dependence

The discussion of the problem of rule-following can be traced back to the Kripkean reading of Wittgenstein's remarks on rules and private language in his *Philosophical Investigations* (Kripke 1982).[1] Since the problem of rule-following is discussed extensively in philosophy, we do not intend to give a detailed account of the problem here.[2] Instead, we simply use Pettit's own characterization. Pettit himself states what the problem of rule-following is about in the following quote:

> "How is basic rule-following possible? How is it possible for a simple creature like you or me, starting from a point that is free of normative connotations, to target a rule in the required sense? How is it possible for us, as purely naturalistic systems, to break into the space of rules and reasons in the first place? This is a basic challenge for anyone who thinks that reasoning is important to mental life." (JD, 243)

Also, Pettit tells us about the motivation of the problem:

> The rule-following problem is often motivated by the fact that everything is like everything in some respect, and that from any finite set of samples there will be nothing inherently wrong about extending the set in any of an indefinite range of directions. (JD, 246)

[1] Many Wittgenstein-commentators are skeptical that Kripke's reading is a correct *interpretation* of Wittgenstein (see e.g. McGinn (1997), chapter 3). Kripke himself is agnostic about this issue (see Kripke (1982), 5).

[2] For an overview see McGinn (1997), chapter 3 and Miller (2002). For a collection of seminal papers see: Miller/Wright (2002).

A. Bahr (✉)
Philosophisches Seminar, Westfälische Wilhelms-Universität, Münster, Germany
e-mail: amrei.bahr@uni-muenster.de

M. Seidel
Zentrum für Wissenschaftstheorie, Westfälische Wilhelms-Universität, Münster, Germany
e-mail: markus.seidel@wwu.de

S. Derpmann, D.P. Schweikard (eds.), *Philip Pettit: Five Themes from his Work*,
Münster Lectures in Philosophy 1, DOI 10.1007/978-3-319-26103-4_2

To illustrate what the problem is about, imagine that you master to apply a certain symbol correctly, say, the well-known symbol "+". If so, you are in a position to apply the symbol correctly in an indefinite number of cases. If you are a competent user of the symbol "+", you follow a certain rule; namely the rule of addition. But how can a finite set of examples (like that set of examples in which you correctly operated with "+") determine just the one rule of addition? What determines the correct way of using the symbol in a new situation? Who is to say that by "+" I did not mean another operation – call it "quaddition" – according to which it is correct to answer "68 + 57" by "5"?[3] How is it possible to avoid a skeptical conclusion about rule-following and correctness? (Kripke (1982), 21)

The challenge posed by the problem of rule-following can – as Pettit himself believes – hardly be overstated: "There is no extant philosophical challenge that compares on the scale of iconoclasm with the skeptical challenge to rule-following" (Pettit 2002b, 31). If we do not find a way to reject skepticism about rule-following, then we "put in jeopardy some of our most central notions about ourselves" (ibid.). In the end, the basic challenge of the problem of rule-following is to provide an account of how we can target normativity and correctness.

In order to understand how Pettit aims to avoid the skeptical conclusion, one should consider two prominent, competing attempts to solve the rule-following problem: the dispositional solution and the communitarian solution. The basic idea of the dispositionalist is that "[t]he claim that [by '+'] I meant *addition* would be true if I was disposed in the past, when asked to compute 'x + y' to produce the *sum* of the two numbers" (Miller 2002, 8). Kripke – and Pettit seems to agree here (Pettit 2002b, 34f.) – argues that the dispositionalist fails to account for the normativity of meaning: "The point is *not* that, if I meant addition by '+', I *will* answer '125', but that, if I intend to accord with my past meaning of '+', I *should* answer '125'. […] The relation of meaning and intention to future action is *normative*, not *descriptive*" (Kripke 1982, 37). The communitarian account – sometimes attributed to Kripke himself[4] – suggests that "what makes your answer '125' correct and the answer '5' incorrect is that others in your community agree with your results" (Kusch 2002, 210).[5] This solution, however, appears to lead to a form of general relativism concerning normativity which is unpalatable for many philosophers. We will see in the following that Pettit's account can be described to aim at incorporating the best aspects of both accounts: dispositionalist and communiatrian – obviously by trying to do so without buying in the problems.[6]

[3] Our example is, of course, Kripke's famous example of "quus": "So perhaps in the past I used 'plus' and '+' to denote a function which I will call 'quus' and symbolize by '⊕'. It is defined by: $x \oplus y = x + y$, if $x,y < 57$, [or: $x \oplus y = 5$, if $x,y \geq 57$]. Who is to say that this is not the function I previously meant by '+'?" (Kripke (1982), 8f.).

[4] See Boghossian (2002), 157, Kusch (2002), 210.

[5] See also Wright (1980), 219–220.

[6] In response to our talk Pettit admitted that his account can be described to be a kind of mixture of these two classical accounts.

The key to Pettit's attempt of solving the rule-following problem is global response-dependence.[7] We can describe his proposed solution by summarizing it in three steps (see Pettit 2002a, 4ff).

[First Step:] For a given subject, "a finite set of examples may exemplify a more or less determinate way of going on" (Pettit 2002a, 4). That means that "although any finite set of examples instantiates an indefinite number of rules, for a particular agent the set may exemplify just one rule" (Pettit 2002b, 36). If so, then it is possible that [Second Step:] "the examples induce a more or less blind disposition to extrapolate in that direction" (Pettit 2002a, 4).

Up to this point, we have a classical dispositional answer to the problem; an answer that Pettit himself rejects (see Pettit 2002b, 34; CM, 85/86). The first and second step are not sufficient to solve the problem of rule-following since agents cannot be mistaken if they only possess this *individual* disposition: Pettit insists that the relationship between the disposition and the rule needs to be explained. As Pettit says, steps 1 & 2 provide us with an account of the independent identifiability and the direct readability of rules, but the fallible readability is excluded:

> Under the story we have developed, a certain sort of biconditional holds a priori: [...] It is going to be a priori under the story so far that a new case x is an instance of the rule r if and only if S is disposed, given full information about x, to treat it as similar in relevant respects to cases that it takes to exemplify r. In other words, according to our story, there is going to be an a priori connection between the rule r and the inclination in virtue of which certain cases exemplify the rule (CM, 91).

However, according to Pettit, "[i]f the rule is a priori connected in this way with the subject's inclination, then there is no room for fallibility" (CM, 91). How is it possible that the inclination can mislead the subject on some occasions? Pettit argues that we must distinguish between favorable and unfavorable circumstances in which the inclination is operative. If we can do so, then "the inclination in itself will not be a priori connected with the rule; it will be a priori connected with it in favorable circumstances only. And so there will be a possibility of securing fallibility" (CM, 92).

We can see now why Pettit thinks that global response-dependence is necessary to solve the rule-following problem: The a priori biconditionals are introduced to illuminate the relationship between our inclinations to go on in a certain way and the *correct* way to go on in a certain way. But nevertheless, they should allow for the possibility of our inclination to go wrong, that is, in cases where circumstances are not favorable. But which circumstances are favorable or normal? Pettit refers to an account he calls "ethocentric" to answer this question:

> [Third Step:] "Normal circumstances are identified by reference to how the participants carry on; they are the circumstances that survive intrapersonal or interpersonal negotiation about how to handle discrepant responses. [...] The account is ethocentric, because it identifies normal circumstances by reference, first, to the habits of response among subjects –

[7] "And I have argued that, if we are to make sense of thinking, in particular if we are to resolve Kripke's version of the Wittgensteinian problem of rule-following, then we must acknowledge a global form of response-dependence." (Pettit (2002c), 50).

> say, their dispositions to have certain sensations in the presence of red objects – and, second, to their practices of negotiation about discrepancies in those responses." (CM, 93).

Therefore, since on Pettit's ethocentric account normal circumstances are identified with habits of and negotiations over *responses*, it implies a global form of so-called "response-dependence". This, in fact, is the communitarian element in Pettit's dispositional account. Pettit sums up:

> If a rule is to be fallibly readable, then it cannot be a priori connected with the inclination in virtue of which it is exemplified and under the influence of which it is read. It must be connected to that inclination in an a posteriori fashion, so that it is a matter for empirical checking that the inclination leads a subject correctly or incorrectly. What can the connection be? In particular, what can the connection be, given that we have to keep enough of our earlier story in place to explain the independent identifiability and the direct readability of rules? The answer to which we have been driven, and there is no obvious alternative, is that the rule is a priori connected to the inclination under circumstances that count, by the ethocentric criterion, as normal or ideal (CM, 95f.).

Therefore, with Pettit's ethocentric account in terms of habits of and negotiations over responses, we can see that according to Pettit, global response-dependence is central to solving the problem of rule-following. Unfortunately, we have to admit, that we have not yet understood what global response-dependence exactly is about, and in turn, we therefore do not see how it solves the rule-following problem without being committed to implausible consequences.

2.2 What Is Response-Dependence?

In order to get a grip on response-dependence, we should note that for Pettit a certain a priori biconditional is decisive: "When a basic term or concept is response-dependent [...] that will entitle us to assert a certain *a priori* biconditional." (Jackson and Pettit 2002, 100).

Our puzzlement with respect to response-dependence stems from Pettit's formulation(s) of the response-dependence biconditional(s) and has mainly two sources. First, we do not understand what exactly the so-called "response-dependent biconditonal" is supposed to connect. Second, we are not really sure what, according to Pettit, "a priori" means in the context of the response-dependent biconditional. In this section, we will address these two issues separately.

2.2.1 What Is the Response-Dependent Biconditional?

In his most recent statement on the issue in question, Pettit distances himself from some of his earlier formulations of the response-dependent biconditional, suggesting that he has changed his position on response-dependence (see JD, 248, Fn. 19). A motivation for that could have been the harsh criticism of response-dependence by authors from the realist camp; these authors believe that global response-dependence "provides an example of the semantic path to Worldmaking and for that reason alone should be rejected" (Devitt 2006, 3).[8] We think that in reaction to this, Pettit aims to move to more cautious formulations of response-dependence than in his earlier papers. However, we do not think that Pettit was completely uncautious before and is completely cautious now, but that he has been and still is cautious and uncautious in all of his papers. Thus in early papers, we find different formulations of response-dependence within one and the same text, and alas, this situation has not completely changed until today.

To prove that, let us take a look at several quotes from different papers of Pettit from 1991 to 2007; we will always compare the quotes within one paper to show that the formulations differ not only from paper to paper, but also in one paper. We will start with one of his earliest statements on response-dependence.

2.2.1.1 Realism and Response-Dependence (1991)

The first formulation we want to have a look at is the following:

> Given our story about the concept of red, we can see how it can come to be a priori knowable that something is red if and only if it is such as to look red to normal observers in normal circumstances (Pettit 2002c, 66).

This suggests that something has a property T iff it seems T to normal observers in normal circumstances. However, one page later, Pettit's claim appears to be slightly different:

> [My story] points out that we commentators are in a position to hold it to be a priori that something is red for the participants in a discourse if and only if it looks red to them under conditions that survive negotiation across times and persons: that is, under conditions that count as normal (Pettit 2002c, 67).

Here, compared with the quote before, we have a constraint: In this quote, we seem to have the idea that something has the property T relative to a certain group of speakers. As far as we are concerned, this is different from the claim in the quote before. Nevertheless, what is connected biconditionally in both quotes are properties and sensations. However, one page later again, Pettit states:

> Consider the biconditional for the concept of redness. [...] With [...] a biconditional understood in the light of our ethocentric story [...] we are presented with concepts on the right-hand side such that it is not so much a grasp of those concepts, but rather a capacity to display the responses and follow the practices to which the concepts refer us, that yields all that is required for a proper grasp of the target concept (Pettit 2002c, 70).

We are puzzled. Why does Pettit claim here that the response-dependent biconditional connects (the concept of) redness with other concepts? Before, we always had sensations or dispositions and not concepts on the right-hand side of the biconditional. Furthermore and relatedly, we are puzzled with the way Pettit denotes the

[8] For a discussion of realism and response-dependence see also Norris (2005).

response-dependent biconditional: In the latter quote, he claims that it is a biconditional for the concept of redness. This biconditional is also called shortly "*the* response-dependence biconditional" (Pettit 2002c, 71, emphasis added) by Pettit on the next page. However, at the top of page 70, he speaks of an "[…] a priori biconditional associated with redness – *the* response-dependence biconditional – […]" (Pettit 2002c, 69/70, emphasis added). Thus it seems to us that sometimes, *the* response-dependence biconditional is about the *concept of redness*, and sometimes it seems to be about *redness itself*.[9]

2.2.1.2 Terms, Things and Response-Dependence (1998)

We now turn to a later paper by Pettit. In his *Terms, Things, Response-Dependence,* Pettit introduces the response-dependent biconditional in the following way:

> Colour terms provide the least contentious examples, for it is agreed on many sides that something is red, as an *a priori* matter, just in case it is disposed to look red to normal observers in normal circumstances. (Pettit 1998, 55)[10]

The quote seems to state that it is a priori that something has the property T iff it is disposed to seem T to normal observers in normal circumstances. We should be careful not to overstate Pettit's claim here: He insists that adhering to such a response-dependent biconditional does not imply that redness is a dispositional property.[11] Nevertheless,

> [e]ven if we do not think about redness as dispositional, that property does indeed have a dispositional aspect: if they can name the property, then its presence in something goes *a priori* with the thing's being disposed to look red to normal observers in normal conditions. (Pettit 1998, 58).

This quote expresses the idea that if a property T can be named, it is a priori that something is T iff it is disposed to seem T. Note that this is different from the former formulation of the response-dependent biconditional: Pettit mentions a condition for the correctness of the a priori biconditional, namely the denominability of the property in question. The addition of this condition of denominability is by itself unproblematic, because it does not change the a priori biconditional as it was stated before. However, let us have a look at the following quote: "Why is it *a priori*, then, that if they can name the property, something is red if and only if it looks red to normal observers in normal conditions?" (Pettit 1998, 62).

[9] "It does not say that something is red if and only if it looks red in conditions that ensure that red things look red; it says that something is red if and only if it answers in a certain way to the sensations and practices of those who use the concept." (Pettit (2002c), 68).

[10] See also: "The claim that it is *a priori* that something is red just in case it is disposed to look red to normal observers in normal circumstances" (Pettit (1998), 56).

[11] "Under these accounts we do not think of the property of redness as a disposition, in the way that we think of fragility as a disposition, and yet the *a priori* linkage between redness and the disposition to produce certain sensations of redness is firmly established" (Pettit (1998), 56).

Here, we have the impression that something else is a priori[12]: It is not just the biconditional, but a new conditional whose antecedent is the denominability of T and whose consequent is the biconditional. Thus, we do not have the statement that if something is denominably T, then it is a priori that it is T iff it is disposed to seem T, but the statement that it is a priori that if something is denominably T, then it is T iff it is disposed to seem T. Expressed in a semi-formal way we have the statement "Denominably T → A Priori (T ↔ Disposed to look T)" in contrast to the rather different statement "A Priori (Denominably T → (T ↔ Disposed to look T))". We are puzzled about the scope of the a priori here.

Though we have a problem with the scope, until now the biconditional itself has not changed: It was always the property of T that was biconditionally connected to the disposition to seem T. However, let us have a look at the following quote: "And so it is *a priori* – it is knowable from a knowledge of how we are guided – that a certain mind-independent property will count as redness if and only if it engenders appropriate looks." (Pettit 1998, 62)

Here, it is a priori that a property will count as T iff it seems T. It seems to us that this is quite different from the former formulations: In all other formulations, it is a priori that something is T iff it seems T. We take it that there is much difference, especially once it concerns the question of realism whether we make a statement about something being T or something counting as T.[13]

2.2.1.3 Overview (2002)

We want to focus our attention to a later part of Pettit's work now. In the *Overview* of a collection of papers on *Rules, Reasons, and Norms* he summarized his proposed solution to the rule-following problem. Here he characterizes response-dependence thus:

> This response-dependence means that the representations – the terms or concepts – mastered satisfy a certain *a priori* biconditional. […] In schematic form, it implies that, for a given term or concept T, it is *a priori* that something is T if and only if it is such as to seem T in favorable conditions (Pettit 2002a, 11).

[12] Note that there is also another difference between the quotes: in the first one, Pettit speaks about a *disposition* to look red whereas in the second quote he talks about looks. We are not sure whether this makes any difference according to Pettit.

[13] There are some other quotes we do not really understand in *Terms, Things and Response-dependence*, e.g. the following one: "Consistently with 'red' referring to a mind-independent property, […] there will still be a question as to why that property gets to be identified as redness. And the answer to that question must be that given how we use the word 'red', it is *a priori* knowable that it will refer to a property that causes things to look red" (Pettit (1998), 62). Here, we have a priori knowledge not of a biconditional, but of some causal connection. We leave it to the reader to decide whether this is a new formulation of the biconditional or not.

In this quote, we find the idea – to take Pettit's example of redness – that for the term "red", it is a priori that something is red iff it is such as to seem red.[14] Note that what is connected here in an a priori manner is the property of being red with the sensation of redness in favorable conditions. Expressed semiformally, we can say that according to this formulation, the following holds true:

(a) A priori (property of T ↔ sensation of T)

This impression is confirmed, though in a slightly different manner, by the following quote four pages later:

> [I]t is not in virtue of the nature of redness that there is an *a priori* connection between the property and such red sensation. Rather, it is in virtue of the denominability of redness, as pinted ot earlier, that such a connection obtains (Pettit 2002a, 15f.).

Again, we have an a priori connection between a certain property and a certain sensation. The difference is that in the latter quote, Pettit provides us with a reason or condition for this connection to obtain, namely the denominability of the property in question. As Pettit explains,

> It is *a priori* that red things are those that actually look red in favorable conditions because we identify the property that we ascribe with the predicate 'red' – we fix the semantic value of the predicate – by the fact that it produces such sensations (Pettit 2002a, 16).

Expressed semiformally, the idea seems to be the following: Denominability of T → A priori (property of T ↔ sensation of T). In any case, though Pettit speaks of the denominability of T, what is connected a priori still is a property with a sensation. It is surprising, however, that Pettit gives a different formulation only two pages before the just quoted passage:

> What is linked *a priori* with being seen as T in favorable conditions, then, is not so much the fact of something's being T as the related fact of its deserving to be designated and treated as 'T': the fact of being denominably T (Pettit 2002a, 13).

We see, that also in this case, the denominability of T plays a crucial role. However, we think it is obvious that this formulation does not connect a property T with the sensation of T because of the denominability of T, but that in this quote we find the claim to an a priori connection between the denominability of T and the sensation of T. Most importantly, Pettit seems to reject that response-dependence implies an a priori connection between a property T and the sensation of T, which he endorses in the other quotes we have mentioned. Therefore, it seems that Pettit wants to express the following two biconditionals:

(b) Not (A Priori (property of T ↔ sensation of T))
(c) A Priori (Denominability of T ↔ sensation of T).

[14] "It is *a priori* that something is red, in the canonical example, if and only if it is such as to look red in favorable conditions" (Pettit (2002a), 11); "On the account developed here being red is connected *a priori* with a certain response: looking red in favorable conditions […]" (ibid., 15).

As far as we can see, these statements are at best different from and at worst contradictory to his other statements in *Overview* – most obviously to (a).[15]

2.2.1.4 Joining the Dots (2007)

Pettit might simply object: "You're completely right. This was all muddle-headed. But you have not noticed that in my recent 'Joining the dots' I explicitly accommodate your concern and move to a more cautious formulation of the biconditional. And that is my official position now. So you're preaching the converted!" We answer: Admitted. But the puzzle goes on. Here is Pettit's cautious formulation:

> "For any response-dependently mastered term or concept, F, then, it will be a priori knowable that something is available to be named as F if and only if it is such as to give rise to the appropriate response under favorable conditions. It will be necessary and sufficient for the denominability of something as F that it be the sort of thing that occasions the relevant response in favorable conditions. And this will be a priori knowable, being knowable on the basis of considerations to do with how a basic term like 'F' is mastered and gains its meaning." [Footnote 19 attached to the second sentence: "I moved to this cautious formulation, employing the notion of denominability, once it became clear to me that the fact that something is denominable as 'F' only if a certain condition holds does not entail that it is F only if the condition holds"] (JD, 248f.).

We think that also this formulation does not offer a clear statement of the response-dependent biconditional. Let us formulate our criticism by focusing on the following question: Is Pettit talking about *properties* or *concepts* in this quote? At the beginning of the quote Pettit introduces the symbol 'F' to stand for a *term* or a *concept*; i.e. the term – *without quotation marks* – F. However, in formulating his cautious biconditional Pettit appears to use the symbol 'F' to denote *properties*. This is most obvious from the fact that Pettit also speaks of the term – *in quotation marks* – 'F'; thus, he switches between talking about the term F and about the term 'F' and between talking of the denominability of something as F and the denominability of something as 'F'. In view of the fact that Pettit's explicit intention for introducing his cautious formulation is his acknowledgement of the difference between talking about something being denominably 'F' and something being F, it is astonishing that he does not seem to adhere to this distinction consistently in his own formulation.

[15] In the paper *Response-Dependence without Tears* that was also published in 2002 and co-authored with Frank Jackson, we seem to have the same incongruity. Thus, we find the following statement: "It may be *a priori* that something is denominably T, as we can put it, if and only if it is such as to seem T under independent, favorable specifications." (Jackson and Pettit (2002a), 102). Here, we have an a priori connection between something being *denominably* t and something seeming T. On the same page, however, it is also referred to the "*a priori* connection between being and seeming, […] a connection that supposes denominability" (ibid.). Here, denominability is a *condition for* the connection between *being* and seeming. We think, as the title of our paper suggests, that we are far away from response-dependence *without tears*.

2.2.1.5 Summary

To sum up, we think that Pettit's characterization of response-dependence is not totally clear: We have seen that in papers from 1991, 1998, 2002 and 2007, we find similar ambiguities. First of all, it is not quite clear what exactly is supposed to be a priori; the scope of the a priori changes in the different formulations of the response-dependence biconditional. Though this is unfortunate with respect to clarity and precludes an understanding of response-dependence, by itself this fact is not necessarily philosophically questionable. As opposed to this, the other ambiguity we have found could have wide-ranging philosophical consequences; it seems to us that Pettit's biconditionals sometimes connect properties and sensations (or dispositions), sometimes terms and sensations (or dispositions) and sometimes concepts with other concepts. We approvingly quote Pettit on this matter: In order to defend global response-dependence against charges of wholehearted realists, Pettit insists that

> [t]he crucial point to grasp is that there is a large difference between saying that the non-parasitic mastery of a predicate is response-dependent and saying that the predicate itself is definitionally response-dependent or, even more extremely, that the property it ascribes is ontologically response-dependent. (JD, 247)

We completely agree with this, as any realist will do. But we have seen that Pettit himself defines response-dependence in all the supposedly different ways he mentions in the quote, so that it is entirely difficult to grasp what response-dependence is supposed to be, and consequently, the implications of response-dependence for the question of realism remain unclear. Pettit himself thinks that response-dependence is consistent with realism[16]; however, we must admit we are not even sure about that.[17]

2.2.2 *What Does "A Priori" Mean in the Context of the Response-Dependent Biconditional?*

"Here, and henceforth, the notion of the a priori is introduced without a commitment to any particular theory." (Pettit 2002c, 52, Footnote 8; same note in CM, 107, endnote 7).

Pettit talks a lot about a priori in the course of discussing response-dependence, although in this context, as he states, he does not commit himself to any particular theory. We agree with Pettit that in this context it is not required to have a detailed theory of the a priori; nevertheless, we believe that for understanding response-dependence, it would be very helpful to know precisely what Pettit means

[16] Cf. JD, 248; Menzies and Pettit (1993), 100; Pettit (2002c), 75.

[17] Thus, we find the statement that "response-dependence fits well with realism, even with the cosmocentric aspect of realism" (Pettit (2002a), 14) and also the statement that "response-dependence entails a rejection of the realist cosmocentric thesis" (ibid., 17). We do not see how *the same* thesis can both fit well with *as well as* entail a rejection of the cosmocentric aspect of realism.

by "a priori". Moreover, whether or not there is any a priori knowledge is highly disputed (Devitt 1998). A necessary condition for deciding whether there is any a priori knowledge would be to know exactly what it means to say that something is a priori. This is important also because many believe that the notion of the a priori is not coextensive with concepts like necessity, infallibility and analyticity.[18]

In order to find out what Pettit means by "a priori", let us start with the following quote:

> [T]his [...] biconditional is true *a priori*. Anyone who is party to the way people follow their sensations and adjust in face of discrepancies will be in a position to know the truth of the biconditional; it does not require empirical information (Pettit 1998, 58).

This quote seems to contain two characterizations of the a priori; we are not quite sure whether these are coextensive or not. The second one seems to be the classic conception of the a priori (see Kant 1998, B2), which can be found also in other places in Pettit's work. Thus in *The Common Mind*, he states the following: "Under the story we have developed, a certain sort of biconditional holds a priori: it holds in such a way that we do not have to employ ordinary empirical checks to establish its truth" (CM, 91).[19] It remains unclear, however, how the first characterization in the first quote relates to the second characterization in the first quote. Is it really true that we do not need empirical information in order to be party to the way people follow their sensations and adjust in face of discrepancies? What exactly is required to become such a competent follower of sensations?

A few pages after the first quote, Pettit claims with respect to the response-dependent biconditional that "it is *a priori* – it is knowable from a knowledge of how we are guided" (Pettit 1998, 62). Pettit probably means the same here as he meant in the first part of the first quote. However, we would appreciate if Pettit could elaborate on this notion of a priori since it is at least not obvious that this characterization is coextensive with the classical one. We suspect that what Pettit has in mind here is a specific thesis about language usage. This is also the case in the following quotes, though in a slightly different manner: "[...] and this is necessarily knowable to anyone who understands the utterances; it is knowable a priori" (Pettit 2002c, 52).[20]

In these formulations it is clear that Pettit takes a priori knowledge to be knowledge on the basis of linguistic competence. Thus the notion of a priori does not only

[18] See e.g. Kripke (1980), Scholz (2009), and Spohn (2009), 31–36.

[19] See also Pettit's characterization of a posteriori: "It must be connected to that inclination in an a posteriori fashion, so that it is a matter for empirical checking that the inclination leads a subject correctly or incorrectly" (CM, 95/96).

[20] See also: "And this will be a priori knowable, being knowable on the basis of considerations to do with how a basic term like 'F' is mastered and gains its meaning." (JD, 249); "It is a priori knowable, it is knowable just in virtue of understanding how the referent of the concept is fixed." (CM, 196). "But what the biconditional tells us is still plausibly a priori. Knowledge of the practices current among those who use the concept is sufficient to give knowledge of the truth of the proposition; we do not have to know in detail about which conditions actually pass the discounting test." (Pettit (2002c), 68/69).

seem to denote an epistemic category, as was suggested by the classical formulation, but also a category having to do with linguistic considerations. Note that in philosophical discussions, usually the a priori/a posteriori distinction is understood as an epistemic distinction, whereas once it concerns knowledge on the basis of linguistic competence, it is the analytic/synthetic-distinction that is used. We are not altogether sure whether Pettit wants to differentiate between apriority and analyticity. This is especially important because Pettit at one point claims that the response-dependent biconditionals express "a priori, contingent truth[s] on a par with the a priori synthetic truths to which Kant gave such importance"(Pettit 2002c, 90). Just as we are not sure about the difference/connection between apriority and analyticity, we are also not sure if Pettit wants to differentiate between apriority and infallibility; the following quote could point to the fact that he does not want that: "On such a view it is a priori that the participants are correct in a large number of their claims: thus there are limits on error, and anthropocentrism holds" (Pettit 2002c, 56).[21]

We would like to know if Pettit thinks that a priori knowledge implies infallibility, especially because recent discussion has suggested the contrary. (see Scholz 2009 passim, Spohn 2009, 33).

2.3 Conclusion

We tried to elucidate Pettit's proposed solution to the very important philosophical problem of rule-following. Pettit wants to establish an account of global response-dependence in order to free us from rule-following skepticism. We scrutinized this account by focusing on his a priori response-dependent biconditionals. Our discussion has shown that Pettit's proposed solution is not totally clear because he does not seem to differentiate between quite different formulations of the response-dependent biconditional. As far as we see, this unclarity is debilitating with respect to a final assessment of whether global response-dependence is consistent with realism. Though Pettit thinks that we can have response-dependence without tears, we are driven to the conclusion that by accepting global response-dependence we are left drowning in a sea of tears. Can Pettit drain this sea by conclusively answering the following questions?

(a) What exactly is supposed to be a priori? How can the change of scope of the a priori in the different formulations be explained?
(b) What exactly is the response-dependent biconditional supposed to connect?
(c) What is the relationship between the a priori/a posteriori distinction and the analytic/synthetic distinction?
(d) What is the relationship between apriority and infallibility?

[21] See also Pettit (2002c), 69.

References

Boghossian, P.A. 2002. *The rule-following considerations*. In Miller/Wright. 2002, 141–187.

Devitt, M. 1998. Naturalism and the A Priori. *Philosophical Studies* 92: 45–65.

Devitt, M. 2006. Worldmaking made hard: Rejecting global response dependency. *Croatian Journal of Philosophy* 16: 3–25.

Jackson, F., and P. Pettit. 2002. Response-dependence without tears. *Philosophical Issues* 12: 97–117.

Kant, I. 1998. *Kritik der reinen Vernunft*. Hamburg: Meiner.

Kripke, S.A. 1980. *Naming and necessity*. Cambridge: Harvard University Press.

Kripke, S.A. 1982. *Wittgenstein on rules and private language*. Oxford: Blackwell.

Kusch, M. 2002. *Knowledge by agreement. The programme of communitarian epistemology*. Oxford: Clarendon.

McGinn, M. 1997. *Wittgenstein and the philosophical investigations*. London: Routledge.

Menzies, P., and P. Pettit. 1993. Found: The missing explanation. *Analysis* 53: 100–109.

Miller, A. 2002. *Introduction*. In Miller/Wright. 2002, 1–15.

Miller, A., and C. Wright (eds.). 2002. *Rule-following and meaning*. Chesham: Acumen.

Norris, C. 2005. *Truth matters: Realism, anti-realism and response-dependence*. Edinburgh: Edinburgh University Press.

Pettit, P. 1998. Terms, things and response-dependence. *European Review of Philosophy* 3: 55–66.

Pettit, P. 2002a. *Overview. In his: Rules, reasons and norms: Selected essays*, 3–25. Oxford: Oxford University Press.

Pettit, P. 2002b. *The reality of rule-following. In his: Rules, reasons and norms: Selected essays*, 26–48. Oxford: Oxford University Press.

Pettit, P. 2002c. *Realism and response-dependence. In his: Rules, reasons and norms: Selected essays*, 49–95. Oxford: Oxford University Press.

Scholz, O.R. 2009. The methodology of presumption rules – Between the A Priori and the A Posteriori. In *The A Priori and its role in philosophy*, ed. N. Kompa, C. Nimtz, and C. Suhm, 173–184. Paderborn: Mentis.

Spohn, W. 2009. A Priori reasons: Two difficult notions and an even more difficult connection. In *The A Priori and its role in philosophy*, ed. N. Kompa, C. Nimtz, and C. Suhm, 25–38. Paderborn: Mentis.

Wright, C. 1980. *Wittgenstein on the foundations of mathematics*. Cambridge: Harvard University Press.

Chapter 3
Pettit's Mixed Causal Descriptivism: Feeling Blue

Amrei Bahr, Bianca Hüsing, and Jan G. Michel

3.1 Introduction

According to traditional descriptivism in the philosophy of language, competent speakers of a given language associate definite descriptions with a proper name "n" and use "n" with the intention to refer to whatever it is that satisfies these descriptions. However, traditional descriptivism has taken a lot of flak for many reasons in the last decades, especially on the grounds of the powerful modal, epistemological and semantic arguments that Saul Kripke provided (Kripke 1980). Kripke's arguments lead to the conclusions that descriptivism is false and that proper names do not have descriptive or informative contents, but are rigid designators. Rigid designators can be characterized as follows: a term is a rigid designator iff it refers to the same entity in all possible worlds in which the entity exists and never refers to anything else.

To illustrate, let us take the classic example[1]: suppose that speakers associate the definite description "the last great philosopher of antiquity" with the proper name "Aristotle". On Kripke's analysis, the proper name "Aristotle" is a rigid designator, while the definite description "the last great philosopher of antiquity" is a non-rigid designator. That in turn means that "Aristotle" refers to the same person in all possible worlds in which the person exists, while "the last great philosopher of antiquity" does not refer to the same person in all possible worlds.

Kripke's important insight is that there is an unbridgeable gap between rigid and non-rigid designators: do you know what the rigid proper name "Adrastea" refers to? If you do not know it yet, you will not find it out by investigating the name alone, i.e., by conceptual analysis, because the name does not contain any semantic

[1] This is the example Frege gave in 1892 already (cf. 2002, fn. 2). Pettit gives it, too, cf. DRA, 324f.

A. Bahr • B. Hüsing • J.G. Michel (✉)
Philosophisches Seminar, Westfälische Wilhelms-Universität, Münster, Germany
e-mail: amrei.bahr@uni-muenster.de; bianca.huesing@gmail.com; jagumi@gmx.de

S. Derpmann, D.P. Schweikard (eds.), *Philip Pettit: Five Themes from his Work*,
Münster Lectures in Philosophy 1, DOI 10.1007/978-3-319-26103-4_3

information that may be of any help. And this is an essential feature of rigid designators: their only semantic function is to refer, they have no informative content over and above that. By way of contrast, what makes definite descriptions such as "the second-closest moon to Jupiter" non-rigid is the informative element they contain over and above their reference function. That means: even if you do not know what the referent of the expression "the second-closest moon to Jupiter" is, you can find it out by the information given.

To make a long story short, this debate, as Pettit puts it, "has prompted the appearance of a new variant of the doctrine [of descriptivism], sometimes cast as rigidified descriptivism" (DRA, 324). Rigidified descriptivists, e.g., David Chalmers and Frank Jackson, are driven by the hope that it is possible to rigidify non-rigid definite descriptions and, thereby, to overcome the arguments of Kripke and others. In order to rigidify a non-rigid designator, you can, e.g., use an "actuality-index" (DRA, 325), and turn the non-rigid designator "the second-closest moon to Jupiter" into the rigid designator "the actual second-closest moon to Jupiter". Since, in general, indexicals such as "I", "here", "now" or "that" can be used to rigidify non-rigid designators, the world-indexical "actual" in our example rigidifies the non-rigid designator "the second-closest moon to Jupiter".[2] Thus, rigidified descriptivism seems to be a tenable view. Moreover, the formal framework of 2D-semantics seems to support that rigidified descriptivism, or, at least, to provide a useful basis.

However, Robert Stalnaker has argued that rigidified descriptivism is unattractive in several ways. Philip Pettit, in turn, thinks that Stalnaker's "objections can be avoided under a global descriptivism that is constrained in a different manner from Lewis's" (DRA, 333).

It is our aim in this paper to critically examine Pettit's proposed solution. We proceed as follows: in a first step, we take a look at (Pettit's presentation of) Stalnaker's objections against rigidified descriptivism. In a second step, we show what Pettit's proposed solution consists in—he talks of a "story in ten stages" (DRA, 334). In a third step, we finally critically examine Pettit's approach. Since, on the one hand, Pettit illustrates his approach by telling a story about the term "blue" and, on the other hand, we are not fully satisfied with his approach, the subtitle of our paper is "Feeling Blue".

3.2 Pettit's Presentation of Stalnaker's Objections against RD

Having given a brief account of what is meant by Rigidified Descriptivism (RD), we turn our attention to Stalnaker's objections against theories of this kind. In doing this, we follow Pettit's presentation of Stalnaker's critique. Stalnaker's first objection is directed against the notion that the 2D-framework could serve as a strong

[2] That is what Stalnaker calls "the generalized Kaplan interpretation", Stalnaker 2004, 309ff. Cf. DRA, 326.

argument for RD. According to Stalnaker, the 2D-framework, with the help of which terms can be assigned two different kinds of intensions—i.e., primary and secondary ones—, cannot play an exclusively supportive role in favor of RD. Under a 2D-framework, as Stalnaker puts it, other theories can be represented likewise. The framework and with it the distinction between primary and secondary intensions can also be useful to support Stalnaker's claim that terms and sentences only have secondary intensions and that figuring truth conditions in other possible worlds does not tell us "anything about the semantic assignment that the sentence has here in this world" (DRA, 329). Therefore, the framework can even support a theory conflicting with RD.

Since nothing has been said about inherent problematic issues of RD so far, this first argument does not affect RD in a strong manner. Hence, we will concentrate on Stalnaker's more substantial objections which Pettit himself emphasizes in his paper.

Apart from his claim that no exclusive support for RD can be derived from the 2D-framework, Stalnaker also provides three objections to prove that RD is even "inherently unattractive" (DRA, 330). Before illustrating these objections, we need to introduce the distinction between descriptive and foundational semantics as Stalnaker himself does. A descriptive semantics like RD describes the linkage between terms and items in the world, whereas a foundational semantics deals with the question "why the items assigned have a claim to be assigned to those terms"—a foundational semantics aims at explaining "the connections that the other enterprise [i.e., descriptive semantics] describes" (DRA, 330). When it comes to semantics, the theories sketched so far (RD and Pettit's alternative view) both belong to the descriptive side, and are, hence, incapable of explaining the connection between terms and referents. In order to preserve their descriptive theory on the foundational level, defenders of RD must be global descriptivists, according to Stalnaker.

Global descriptivism (including proper names as well as general terms) comprises the claim that the meaning of all terms is determined by a network of descriptive sentences while speakers "intend to refer to corresponding items" (DRA, 331). The theory outlined briefly is the one Stalnaker's main objections attend to.

3.2.1 The Permutation Problem

The first problem Stalnaker points out is the so-called "permutation problem": if we assume that the meaning of a term is determined by descriptive sentences jointly constituting a coherent network, it seems likely that terms can be permuted without jeopardizing the consistency of the network. Pettit puts the problem as follows:

> If the original assignment has 'big' refer to bigger things, for example, 'small' refer to smaller things, then the permutation might reverse this and, provided it introduced compensating changes elsewhere, still manage to make the networking sentences come out as generally true. (DRA, 331)

For the purpose of our paper, this objection can be put aside since Stalnaker himself has already conceded the possibility of solving the permutation problem by constraining global descriptivism (cf. Stalnaker 2004, 314f). The following two objections, however, seem to be more problematic for a descriptivist position.

3.2.2 The Holism Problem

Following Quine, Stalnaker recognizes the difficulty of filtering the relevant descriptive utterances out of an amount of all the sentences describing a term. Since each speaker can be guided by different descriptive sentences in using a term, the risk of holism and solipsism appears. However, instead of expanding on this "troublesome" (DRA, 332) argument and suggesting possible solutions, Pettit passes over to Stalnaker's third main objection which he estimates to be "more original and more pointed" (DRA, 332).

3.2.3 The Indirectness Problem

The objection which Pettit places the most emphasis on represents the most relevant basis for his attempt at defending a theory that includes descriptivist and causal elements as well. Pettit claims that his solution, which will be illustrated in the next section, can best be characterized as a rigidified and anchored form of descriptivism. With his theory, Pettit tries to offer a solution to Stalnaker's objection that speakers do not "get into direct touch" (DRA, 332) with the referents of their utterances. Given that we derive knowledge from description instead of acquaintance, it is the primary intensions of terms which we as speakers "immediatly grasp" (DRA, 332). Stalnaker argues that we only have indirect access to "the reference conditions and the truth conditions that those primary intensions pick out in the world" (DRA, 332). Hence, we would not talk about individual things and would not be able to "know any individual thing or property in itself but only as that thing or property, whatever it may be in itself, that satisfies a certain description" (DRA, 333).

As we will see in the next section, Pettit claims to provide a solution for at least Stalnaker's indirectness objection, telling a story of concept acquisition which grants direct contact with things in the world.

3.3 Pettit's Story in Ten Stages

On the basis of a sketch of Stalnaker's objections, Pettit claims that these objections can be avoided without having to give up global descriptivism, at least, as long as the global descriptivism in question is constrained, though constrained differently from

Lewis's version of global descriptivism (cf. DRA, 333). But beware: as Pettit himself points out, it is not altogether clear whether the theory he proposes is correctly described as a descriptivist theory or not. Pettit puts it as follows: "The theory may not deserve to be characterized as descriptivism, since it introduces a causal as well as a descriptive element" (DRA, 333). We will consider later on if Pettit's theory can justifiably be labeled descriptivist, but before doing that, we will have to give an outline of the theory as it stands.

According to Pettit, his theory "introduces a causal as well as a descriptive element" (DRA, 333). In addition, Pettit tells us that his theory is an anchored doctrine; we shall see later on what is meant by that. Provisionally, we can call Pettit's theory a rigidified and anchored descriptivism, as he himself does in the title of his paper "Descriptivism, Rigidified and Anchored". To be brief, we will subsequently refer to Pettit's theory with the abbreviation "RAD".

Before sketching RAD in detail, let us take a short look at the accomplishments RAD has according to Pettit. Pettit believes that with RAD, Stalnaker's objections we mentioned above can be avoided (cf. DRA, 333). Pettit tells us that RAD is "a way of dealing with the permutation problem" (DRA, 333), the problem that Stalnaker himself finds less serious because he believes that it can be solved by Lewis's constrained descriptivism. Pettit believes that this problem can equally be solved by RAD. To be more precise, according to Pettit, RAD

> should suggest a response to the Quinean worry that no single set of sentences can ever be selected as those that play the role, in a robust and community-wide way, of guiding speakers in the use of their terms. (DRA, 333)

To be even more precise, RAD "should silence the concern that any terms that are guided in that way will allow us only an indirect sort of contact with their referents" (DRA, 333).

Pettit acquaints us with RAD by telling us, as he puts it, a story in ten stages (cf. DRA, 334) about how a speaker learns the use of the term "blue" (DRA, 333). Pettit believes that this is a story that proponents of rigidified descriptivism such as David Chalmers and Frank Jackson "may find congenial" (DRA, 334). We can, thus, record that he sees himself in the tradition of these theories.

In the beginning of his story in ten stages, Pettit wants us to imagine that he experiences a training by his parents or teachers to learn the correct use of the term "blue": Pettit learns to use the term "blue" "under the causal impact […] of the blueness property" (DRA, 334). The ones who teach Pettit do this by pointing out blue things to him and by contrasting them with things that have other colors (cf. DRA, 334). Pettit emphasizes that the requirement for the training to be successful is that he must be able to pick out the relevant aspect of the things used to teach him, namely the aspect of looking blue under normal conditions (cf. DRA, 334).[3] Furthermore, Pettit tells us that the causal linkage that establishes the blueness

[3] There is a problem lurking behind this formulation, namely the well-known qua problem. We do not have space to deal with this problem in detail here, but it should be mentioned that two of the authors of the present paper, Jan G. Michel and Amrei Bahr, have proposed a solution to the qua problem elsewhere (cf. Bahr, Michel, Voltz 2013).

property as the referent of the term "blue" is, as he puts it, epistemically resonant: this linkage will constitute the belief that the objects Pettit takes to be blue do in fact look blue (cf. DRA, 334). By reflecting on the practice Pettit has established due to his learning process, he concludes "that blueness is a property of things that is associated with their looking blue" (DRA, 334). Also, Pettit believes that beyond this, he will be able to see "that blueness is that property […] which lies at the causal origin of the use of 'blue' on [his] part, and indeed that of [his] fellows" (DRA, 334).

Pettit tells us that the story told so far gives us a hint for evaluating the sentence "'Blueness is that ostensive property that makes things look blue in normal conditions'" (DRA, 334): according to Pettit, under the story told, this sentence is *a priori* true, as it is guaranteed by our practice that the sentence is true; the primary intension of the sentence will be true at every world considered as actual, at least, if a similar practice is present in the world in question (cf. DRA, 335). Note that the sentence contains a definite description that has been rigidified, namely the description "*that ostensive* property that makes things look blue in normal conditions". This definite description is a rigidified one because it contains a rigidifying element that is comprised of the words "that ostensive".

Considering the assumption "that 'blue' is used rigidly and that it refers in the actual world to B1, and in other worlds to different properties" (DRA, 335) that Pettit believes to fit with his story, Pettit concludes that the sentence 'Blueness is B1' will be necessary but *a posteriori,* its secondary intension will be necessary, whereas its first intension will be a contingent one (cf. DRA, 335). According to Pettit, the sentence "'Blueness makes things look blue'" (DRA, 335) will therefore be *a priori* but only contingently true, for "there will be possible worlds where blueness as we understand it—B1—does not make things look blue" (DRA, 335).

In conclusion, Pettit claims that his theory is not exposed to Stalnaker's objection of the world only being accessible in an indirect manner (cf. DRA, 336): given that his story is correct,

> [one] will be able to think of the blueness property as *that* (ostended) property, where the property in question is directly available […] but available only in virtue of making things look blue […]. [One] will not be restricted to thinking of it indirectly as whatever actual property, [one knows] not what, that belongs to things which look blue. (DRA, 336)

Regarding the generalizability of his account, Pettit holds that similar stories can be told for other terms introduced on a causal basis, his theory will at least be applicable to terms of this type (DRA, 336).

As for classifying his own theory, Pettit gives a number of hints: according to his own view, his theory is "something close to descriptivism" (DRA, 336). Insofar as the theory he purports incorporates an actuality-index into the relevant descriptions, "it is in that sense a rigidified descriptivism" (DRA, 336). Since it, additionally, requires speakers to be in causal contact with the referents of the terms in question, it is "in that sense […] an anchored […] doctrine" (DRA, 336). Nevertheless, Pettit concedes that his theory may not deserve to be characterized as descriptivism: "It may be better cast as a mixed doctrine that involves causal as well as descriptivist elements" (DRA, 336). But he still believes that "the doctrine sketched is relatively close to the position espoused by Stalnaker's opponents. And certainly it is close enough to raise some questions about how deep his criticisms go" (DRA, 336).

In the following, Pettit's claim of having provided a theory that escapes Stalnaker's objections will be called into question. Moreover, we will ask how RAD can be classified.

3.4 Critical Examination of Pettit's Proposed Solution

Let us take stock of the discussion so far: the doctrine of traditional descriptivism got under attack by Kripke and others who argued that proper names are rigid designators and not synonymous with definite descriptions. The possibility of rigidifying definite descriptions, however, has prompted a new variant of the doctrine, namely rigidified descriptivism (RD). According to RD, non-rigid definite descriptions can be rigidified by the use of indexicals such as "actual", so that RD seems to provide a tenable descriptivist view. However, Stalnaker argues—and Pettit seems follows him in this (cf. DRA, 330)—that RD is not supported by the 2D-framework. Moreover, Pettit[4] seems to follow Stalnaker with regard to the abovementioned permutation problem (cf. DRA, 332 and 333), and both Pettit and Stalnaker think that the permutation problem can be solved under a constrained global descriptivism.

But, as can be seen from his developing an alternative approach to RD, Pettit does not follow Stalnaker when it comes to the holism problem and the indirectness problem. On the contrary, Pettit wants to avoid these objections with the help of a new variant of a constrained global descriptivism, namely his rigidified and anchored descriptivism (RAD). We take it for granted that RAD solves the permutation problem, but can it solve the other problems as well?

Pettit, at least, is certain that his approach "should suggest a response to the Quinean worry", i.e., to the holism problem, and that "it should silence the concern that any terms that are guided in that way will allow us only an indirect sort of contact with their referents" (DRA, 333), i.e., the indirectness problem. According to the indirectness problem, "speakers don't ever get into direct touch with items in the world" (DRA, 332). A central element of Pettit's RAD is the causal linkage that establishes—in his above-mentioned story—the blueness property (as an item in the world) as the referent of the term "blue"; that is what makes Pettit's approach an anchored doctrine. So, obviously, on Pettit's approach, speakers get into touch with items in the world and, thus, Pettit's approach provides a solution to the indirectness problem.

And what about the holism problem? According to the holism problem, which is "a familiar objection to internalist accounts of meaning" (Stalnaker 2004, 316), every speaker "is likely to be guided by any of the sentences we hold true—hence

[4] And Daniel Stoljar joins him in this.

holism—and that we are guided in a different way from others: hence solipsism" (DRA, 332).

As we have already seen, a central element of Pettit's approach consists in the causal linkage between—in his example—the blueness property and the use of the term "blue". Pettit even expands this externalist and anchored element in his otherwise internalist picture: "Will I be able to see, beyond this, that blueness is that property […] which lies at the causal origin of the use of 'blue' on my part, *and indeed that of my fellows?* I think so" (DRA, 334, our emphasis).

If Pettit is right in thinking so, the holism problem can be solved because there is a common causal basis that every speaker's use of language relies on. But: what is Pettit's argument—why does he "think so"? This remains an open question. However, we think that if he answered it persuasively, he would provide a solution to the holism problem.

Now, it seems as if Pettit's RAD provides not only a solution to the permutation problem, to the holism problem and to the indirectness problem, but also a tenable version of RD, representing "a mixed doctrine that involves causal as well as descriptivist elements" (DRA, 336).

However, from our point of view, a more basic and serious problem for all versions of RD has not been mentioned yet. This problem can be put as follows: as has become obvious, rigidifying is a technical move that turns a non-rigid designator into a rigid designator. But remember: it is an essential feature of rigid designators that their only function is to refer; rigid designators have no descriptive or informative content. Therefore, the non-rigid designator "the second-closest moon to Jupiter" is informative, while the rigid(ified) designator "the actual second-closest moon to Jupiter" is not informative. What does that mean? Let us have a look at the following true sentences:

(1) "The second-closest moon to Jupiter is the second closest-moon to Jupiter."
(2) "The second-closest moon to Jupiter is Adrastea."

Obviously, sentence (1) is knowable *a priori,* while (2) is knowable only *a posteriori.* That's not problematic. But what about the following true sentence?

(3) "The second-closest moon to Jupiter is the actual second closest-moon to Jupiter."

What do you think: is (3) knowable *a priori* or *a posteriori?* Let us take a look at both options one after another.

First option: if (3) is knowable *a priori,* the expression "the second-closest moon to Jupiter" must be synonymous with "the actual second-closest moon to Jupiter". From the definition of "synonymity", it follows: if two expressions are synonymous, they have the same meaning, i.e., the same descriptive or informative content. Therefore, since the non-rigid expression "the second-closest moon to Jupiter" is informative, the expression "the actual second-closest moon to Jupiter" must be informative, too. Since, as follows from the Kripkean arguments, informative expressions are non-rigid, the expression "the actual second-closest moon to Jupiter" is non-rigid as well. All in all, the first option leads to the conclusion that the expression "the actual second-closest moon to Jupiter" still has informative content, but has not been rigidified.

Second option: if (3) is knowable *a posteriori,* the expressions in question are not synonymous. The reason for this is that the expression "the actual second-closest moon to Jupiter" has been rigidified and, therefore, behaves just like the proper name "Adrastea" in sentence (2). All in all, the second option leads to the conclusion that the expression "the actual second-closest moon to Jupiter" has been rigidified, but does not have any meaning or informative content over and above its reference.

Now, what rigidified descriptivists are aiming at is the combination of both options: they want rigidified descriptions to be both rigid and descriptive. But then, they have to hold that sentence (3) is knowable both *a priori* and *a posteriori.* That does not work, so you have to decide for one of the options: the first option leads to (traditional) descriptivism and internalism, the second option leads to a theory of direct reference and externalism.

And what about Pettit's version of RD? Is it better off than "standard" RD? Let us take a closer look: in stage 6 of his story in ten stages, rigidification comes in. Pettit writes: "'Blueness is that ostensive property that makes things look blue in normal conditions'" (DRA, 335). Here, the non-rigid definite description "the property that makes things look blue in normal conditions" is rigidified, but not by using the world-indexical "actual", but by using the context-sensitive indexical "that ostensive".[5] However, the effect is the same: the definite description is rigidified. So, we face the same situation as above, which can be illustrated with the help of the following true sentences:

(4) "Blueness is the property that makes things look blue in normal conditions."
(5) "Blueness is that ostensive property that makes things look blue in normal conditions."

Let us assume that sentence (4) is knowable *a priori.* What about sentence (5)? Is it knowable *a priori* or *a posteriori?* It seems to be on a par with sentence (3) so that we have the two options mentioned above.

First option: if (5) is knowable *a priori,* the expression "blueness" must be synonymous with the expression "that ostensive property that makes things look blue in normal conditions". If two expressions are synonymous, they have the same informative content. Therefore, since the non-rigid expression "blueness" is informative, the expression "that ostensive property that makes things look blue in normal conditions" must be informative, too. Since informative expressions are non-rigid, the expression "that ostensive property that makes things look blue in normal conditions" is non-rigid as well. All in all, the first option leads to the conclusion that the expression "that ostensive property that makes things look blue in normal conditions" still has informative content, but has not been rigidified.

Second option: if (3) is knowable *a posteriori,* the expressions in question are not synonymous. The reason for this is that the expression "that ostensive property that

[5] By the way, "ostensive" in "that ostensive" is redundant.

makes things look blue in normal conditions" has been rigidified and, therefore, behaves just like the proper name "Adrastea" in sentence (2). Summing up, the second option leads to the conclusion that the expression "that ostensive property that makes things look blue in normal conditions" has been rigidified, but does not have any meaning or informative content over and above its reference.

Again: is sentence (5) knowable *a priori* or *a posteriori?* Pettit says: "this will be *a priori* true for me and my fellow-speakers, for it will be guaranteed to be true by the practice that we share with one another" (DRA, 335).

In other words: Pettit chooses the first option. As we have shown above, that, in turn, means that Pettit takes the (traditional) descriptivist and internalist route. And that, in turn, means that Pettit has to deal with Kripke's arguments and Stalnaker's objections once again. Furthermore, since a mixed doctrine seems hardly tenable, Pettit has to decide whether he holds an internalist or an externalist account of meaning.

Finally, it becomes clear why we are feeling blue: Pettit's doctrine of mixed causal descriptivism which seemed to be a promising approach until recently has to face several serious problems now. However, we are eager to see how Pettit deals with these problems.

References

Bahr, A., J.G. Michel, and M. Voltz. 2013. Refining Kitcher's semantics for kind terms, or: Cleaning up the mess. In *Philip Kitcher: Pragmatic naturalism*, ed. M.I. Kaiser and A. Seide, 91–109. Frankfurt: Ontos.

Frege, G. 2002. Über Sinn und Bedeutung. In *Funktion – Begriff – Bedeutung*, ed. Mark Textor, 23–46. Göttingen: Vandenhoeck & Ruprecht. Originally published in 1892.

Kripke, Saul A. 1980. *Naming and necessity*. Cambridge: Harvard University Press.

Stalnaker, Robert. 2004. Assertion revisited: On the interpretation of two-dimensional modal semantics. *Philosophical Studies* 118: 299–322.

Chapter 4
Discovering the Properties of ‘Qualia’ in Pettit’s Theory of Phenomenal Consciousness

Jonas Dessouky and Tobias Peters

4.1 Introduction

Philip Pettit is physicalist, he defines this stance in the following passage: “Physicalists hold that everything in existence is constituted in some way out of physical materials and that all the laws and regularities that hold in the actual world are fixed by physical laws and regularities” (JD, 216).

A problem for physicalists is generated by the so-called ‘qualia’. In this essay we will discuss the term ‘qualia’ first, followed by a brief description of Pettit’s theory of phenomenal consciousness. We want to emphasize that the *word* ‘qualia’ is not important. When we speak about qualia, we mean a concept that has specific properties. Entities with a cluster of these properties are often called qualia in the literature, so we use this word, but you can choose a different word for this cluster of properties. In the third part of this essay we will argue that Pettit’s theory of phenomenal consciousness is indeed suggesting rather than denying if not the existence of qualia but something that shares their fundamental properties. The conclusion of this paper is finally, since something comparable to qualia exists, this, i.e. phenomenal content, destroys a functionalist view in the philosophy of mind.

4.2 Qualia

Pettit argues in his paper that the existence of qualia would indeed be a problem for physicalism:

J. Dessouky (✉) • T. Peters
Philosophisches Seminar, Westfälische Wilhelms-Universität, Münster, Germany
e-mail: j_dess01@uni-muenster.de; tobiaspeters@hotmail.de

S. Derpmann, D.P. Schweikard (eds.), *Philip Pettit: Five Themes from his Work*,
Münster Lectures in Philosophy 1, DOI 10.1007/978-3-319-26103-4_4

> The question raised here is the crux of the issue, as I see it, between physicalism and non-physicalism about consciousness. If qualia are allowed, then it is always going to be logically possible to have a world that is identical to the actual world in physical respects but that lacks qualia or that displays different qualia. The physical cannot fix qualia by the functional organization it incorporates, since qualia are not functionally characterized. So, for all that the physical nature of the world appears to require, there may or may not be qualia present, or there may or may not be this or that pattern of qualia. Physicalism falls if qualia stand. (JD, 262)

The problem Pettit sees here is that if qualia exist there is a problem for physicalism. This is due to the fact that qualia cannot be incorporated into functionalist philosophies of the mind, simply because they are not functionally characterized. Qualia are the "crux of the issue" because physicalism depends on every entity or incident being physically describable. Not so with qualia. We suppose that the short version of the argument can be summarized like this:

T1: If qualia stand, physicalism is false.
T2: If there are no qualia. Physicalism is possibly true.

We think that what is needed at this stage of the argument is an analysis of what qualia are supposed to be. We choose to refer to Daniel Dennett's list of features that are constitutive of qualia. Dennett's characterization is as follows:

> 'Qualia' is an unfamiliar term for something that could not be more familiar to each of us: the ways things seem to us. As is so often the case with philosophical jargon, it is easier to give examples than to give a definition of the term. Look at a glass of milk at sunset; the way it looks to you – the particular, personal, subjective visual quality of the glass of milk is the quale of your visual experience at the moment. The way the milk tastes to you then is another, gustatory quale, and how it sounds to you as you swallow is an auditory quale. These various 'properties of conscious experience' are prime examples of qualia. (Dennett 1988, 381)

According to Dennett, it is not possible to give a particular definition about what qualia really are. Rather they are only accessible to scientific discussion by giving examples of them. Since we cannot really define what qualia are, we are still able to say what the cause of this impossibility of a definition is. If x is only definable by examples then the following features can be ascribed to x (and therefore to qualia). Dennett gives us four of these features. They are:

1. private
2. intrinsic
3. ineffable
4. immediately apprehensive to consciousness. (Dennett 1988)

We do not want to discuss the argumentation by Dennett against qualia. This would be another issue. We only refer to the properties of qualia listed by Dennett to analyze the concept of qualia. We assume Pettit refers to these properties which are not compatible with physicalism. What are these properties exactly?[1]

[1] The description in the next four paragraphs follows Dennett (1988).

Ad 1: That a mental condition is private means that it cannot be shared with others. My visual experience of, say, red is only accessible to myself. I can say that I see (something) red, even I see an intense red. But I cannot share the phenomenal quality itself. If someone sees (something) red, too, we can only guess that he sees something similar to what I see as red. But it remains a guess. My experience of red is so to say "locked" inside myself. So it is a private entity.

Ad 2: "Intrinsic" is meant to say that qualia are not dependent on factors outside myself. Now, this means that the quale itself is something that is depending on me and nothing else. Of course they may be evoked by an object and its light emission, an apple for example. But qualia, the impression, exist independently of these physiological parameters.

Ad 3: The private and intrinsic characteristics of qualia suggest another feature: their ineffability. This can be understood in a way that we cannot make the quale itself a topic of conversation. This means nothing else than that the phenomenal content is at least at some levels of its complexity not translatable into propositional content. Which means our language is outrunned by qualia. Indeed it is not only linguistically outrunned but also logically: Qualia cannot be expressed since only one subject can hear, see, smell etc. it. The answer to the question "What is it like to see red?" would therefore be "It is red." This answer is purely tautological and circular. It demonstrates that linguistically and logically we cannot make qualia objects of (scientific) discourse, although to see red is an experience and it is somehow to be like for us to see red.

Ad 4: Things are immediately apprehensive to consciousness if they are not mediated to us. This means that in the case of phenomenal content, it is absolutely evident that we see this or that. One would never doubt seeing red if he sees red. It is directly given to us qua perception. In the case of hallucinations there might not be an object represented by this red. So we could see an apple, while there really is no apple. We could even know that there is no apple by belief. But nevertheless I know that I see red because it is evidently given to me phenomenally and therefore evidently.

We have listed four features of qualia above: privacy, intrinsicness, ineffability and direct apprehension. Whoever subscribes to the thesis that phenomenal content has these features would admit that there is something at least similar to qualia. For one cannot characterize phenomenal content in this way and still claim that it has nothing in common with qualia, which would lead to difficulties for a functionalist.

4.3 Pettit on Phenomenal Consciousness – A Brief Summary

In his text "Joining the dots" Phillip Pettit gives us a brief description of his theory of phenomenal consciousness. It will be briefly summarized in this paragraph and give us the basis for the claim that phenomenal content in this theory is identical with qualia or at least is characterized by most of the features that are held in account for qualia in the sense described above.

Phillip Pettit's theory of phenomenal consciousness is based on a higher order representational theory. Representational theories of the mind, as Pettit introduces them, claim that phenomenal contents are representations of objects in the world outside ourselves. But according to Pettit we have to make a difference between single order and higher order representations. A first order representation is when there is a simple representation of an object in the creature, meaning that there are mental incidents that are representations of objects of the world outside the subject. This creature however, Pettit says, does not refer to these representations as representations, meaning it is not aware that each representation represents something at all. It might act perfectly according to these representations, following a certain behavioral disposition such as running after it and eat the object. But such a creature is lost in a world of undefined phenomenal incidents. Pettit concludes that therefore it cannot draw its attention toward the object itself, since no such thing as object exists for this creature. No "attention" in this sense is referring to intentionality – a necessary condition for consciousness (JD).

What is additionally needed, according to Pettit, in order to talk about consciousness is a representation of higher order (at least second order). It is only here that the representations are represented (or presented) as such. That means the subject is aware of the fact that what he is seeing, hearing, etc. is representing objects and is therefore able to know that there are such things as objects outside of the subject. So the creature can draw its attention to the objects. This is only possible for representations of a higher order, since one can only reflect objects, internal or external, if one understands the term "object" at all or recognizes the phenomenal content as representations for things outside oneself. This is suggested by Pettit's following sentence: "[…] the way things are will be represented in the creature […]. But it will not be represented *for* the creature, it will not in itself be a potential object of attention" (JD, 256).

Hence, first-order representations are 'represented *in* a creature' – meaning that there are some phenomenal representations in a creature whereas this creature is not aware that these phenomena represent anything. Second-order representations are 'represented *for* a creature' – meaning that these creatures are aware of the representational character of the phenomenal content. Pettit says that it is only possible in this way the creature can turn to the object and not only act conform.

According to Pettit a higher order representation is constituted perceptually, not propositionally. So the presentation of objects in the outside world is due to our belief first but perceptually presented in this way. Representations of higher order are recognized as such. That means we are literally seeing, hearing, etc. that there are objects existing in an outside world that is different from ourselves.

Pettit characterizes perceptual representations as "holistic" (JD, 257) whereas propositional content is "atomistic" (ibid.). The latter gives us exactly one information, the former an indefinite set, just like a digital watch is showing exactly one time information while an analogue watch is giving us the possibility for unlimited fine grained information about the time. Furthermore, Pettit describes perceptual representation to be “sticky” (JD, 258) and somewhat “isolated”. This refers to the fact that propositions are formed according to our perceptions. Pettit notes that to the contrary perceptions cannot be changed by propositions, i.e. our beliefs: As deep as our beliefs might be, they will never change our perception. Pettit emphasizes that perceptual content is relatively autonomous from propositional content. At last Pettit calls perceptual representations “concrete” (ibid.) while propositional representations are “abstract” (ibid.). This combines the holistic character of representations with a new feature: The complexity of the concrete character is far outrunning the possibility of lingual expression.

4.4 Perceptual Content and Qualia

In this section, we will try to set Dennett’s characterization of qualia in relation to Pettit’s theory of phenomenal consciousness. Our claim will be that phenomenal content as described by Pettit and qualia as described by Dennett have more in common than appears at first sight. So we will finally claim that one could discover entities of incidents in Pettit’s theory that have the same properties as qualia.

Pettit states that perceptual representations outrun the propositional representations. We conclude that what is meant here is that not all perceptual contents can be expressed by language. This suggests that they are not translatable into propositions. In other words they cannot be of any language system, that is, they are not expressible. There is thus an inexpressible rest of the perceptual representation.

Since the phenomenal content of perceptual representation is ineffable it is also to be called private. It remains inaccessible to other subjects. If I cannot express mental incidents it remains disconnected from discourse since it cannot be reasonably be talked about other than in circular ways as stated above.

The phenomenal content in Pettit’s representational theory of mind is immediately apprehensive to the subject. The subject is able to perceive the representations unmediated. It is not affected by former propositional belief. What the subject sees it is evidently seeing and is not mediated by other subjects or other instances. The representations are therefore immediately apprehensive for the subject.

Finally the autonomy of the perceptual representations seems to be equivalent to intrinsicness since they exist independently of propositional content. The stickiness of the perceptual content is due to its intrinsic character and again stress out their private character.

We point out that perceptual representation in Pettit’s theory is (1) ineffable, (2) private and finally (3) immediately apprehensive and (4) intrinsic. We claim that features 1–4 give us sufficient reason to prove that perceptual representations here are really not that different from qualia.

4.5 Conclusion

The conclusion of the previous section is that Pettit's characterization of perceptual representation is suggesting the existence of something like qualia. This has consequences for Pettit's claim that "if qualia stand, physicalism falls". This is due to the fact that qualia fall outside the functionalist reckoning and so do phenomena that share central properties with qualia. This type of phenomena can therefore not be incorporated into physicalism simply because they are private, ineffable, intrinsic and immediately apprehensive. Phenomena of this type are simply distracted from scientific discourse due to their properties as mentioned in section I. If we try to build a philosophy of mind and want to hold on to physicalism, e.g. as a variant of functionalism, we need to develop this theory without entities or incidents that have qualian properties. But, as we have shown qualia exist in Pettit's theory, therefore it cannot be a physicalistic theory.

We merge our analysis of qualia and the analysis of Pettit's theory into one conclusion (C2):

T1: If qualia stand, physicalism is false.
T2: Perceptual content is nothing else than qualia.
C1: Qualia exist.
C2: Physicalism is false.

Reference

Dennett, Daniel C. 1988. Quining Qualia. In *Consciousness in contemporary science*, ed. A. Marcel and E. Bisiach, 381–414. New York: Oxford University Press.

Chapter 5
Playing Pong with the Mind? Pettit's Program Model and Mental Causation

Kim Joris Boström, Gordon Leonhard, and Lisa Steinmetz

5.1 Introduction

A notorious problem in the philosophy of mind is the causal relevance of mental properties, which is also known as the *problem of mental causation*. Intuitively, the mental properties should somehow be causally relevant for the physical properties of the body, so that conscious subjects would be capable of actively controlling their body. Mental causation arises in particular on the grounds of *non-reductive physicalism*, because there the mental states are not reduced to physical states. Philip Pettit has dedicated a significant part of his work to physicalism and causality, and he developed a model to clarify the causal structure within a multilevel hierarchy of properties. Pettit holds that the model can also be applied to the case of mental causation as to provide a satisfying solution to its difficulties. In the following, we will first present some relevant concepts of physicalism and multi-level causality in Pettit's work, and we will introduce the problem of mental causation according to its formulation by Jaegwon Kim. Second, we will present Pettit's program model and two instructive examples of application provided by himself. Third, we will raise some concerns about the application of the program model to the case of mental causation, and we will illustrate our concern by a counterexample where the application of the model in our view fails. Lastly, we will propose some ways to deal with these concerns.

K.J. Boström (✉) • G. Leonhard
Philosophisches Seminar, Zentrum für Wissenschaftstheorie,
Westfälische Wilhelms-Universität, Münster, Germany
e-mail: mail@kim-bostroem.de; g_leon02@uni-muenster.de

L. Steinmetz
Institut für Kognitionswissenschaft, Universität Osnabrück, Osnabrück, Germany
e-mail: lisa.e.steinmetz@gmx.de

S. Derpmann, D.P. Schweikard (eds.), *Philip Pettit: Five Themes from his Work*,
Münster Lectures in Philosophy 1, DOI 10.1007/978-3-319-26103-4_5

5.1.1 Causality

Causality is an asymmetric relation between events, so that if two events A and B are causally related, then this is equivalent to saying that A *causes* B. Causation has to be differentiated from logical implication. For example, if Jones is diagnosed of having AIDS then this logically implies that he has been infected with the HI virus. However, the causal relation would show in the opposite direction, as we would rather say that the HIV infection caused Jones to have AIDS. The causal relation is in general not strict, as it might be the case for someone to be HIV positive without actually getting AIDS. Here the differentiation between type and token causality is relevant. The particular case of Jones' being HIV positive is a particular instance, or token, of the general type of "being HIV positive". While "being HIV positive" in general does not strictly lead to having AIDS, for the particular case of Jones the causal relation factually applies. The causal relation between event types is denoted as *type causality*, the causal relation between event tokens is called *token causality*. It should be mentioned that there is still no definite consensus so far as to how events are precisely defined or when the causal relation actually applies.

5.1.2 Physicalism and Mental Causation

In addition to the conceptual difficulties with the notions of events and of causality, there is the problem of mental causation that seems to stand in conflict with the basic premises of physicalism. Pettit provides four claims as the basis of any physicalist theory (Pettit 1995):

1. There are microphysical entities.
2. Microphysical entities constitute everything.
3. There are microphysical regularities.
4. Microphysical regularities govern everything.

These claims are clearly ontological ones: there is a physical world which constitutes and governs our empirical world. Pettit commits himself to the truth of physicalism:

> I accept the truth of physicalism or materialism or naturalism, as it is variously called, and this forms an important part of the background to many of the positions with which I align myself and to many of the views discussed in this volume. Physicalists hold that everything in existence is constituted in some way out of physical materials and that all the laws and regularities that hold in the actual world are fixed by physical laws and regularities (JD, 216).

A standard claim of physicalism in philosophy of mind is that mental states *supervene* on the physical states, and Pettit follows this claim:

> However 'physical' is understood, the account given of physicalism means that a minimal physical duplicate of the actual world—a physical duplicate to which nothing is

> independently added—would be a duplicate *simpliciter*, a duplicate in every respect (Jackson 1998). This is to say that the non-physical character of the actual world is fixed superveniently on the physical. Short of an independent addition, there is no possibility of a change in the non-physical way the world is without a change in its physical makeup; fix the way it is physically, and you will have fixed the way it is in every respect. (JD, 217, inline citation and emphasis as in the original)

Altogether, there is a vertical hierarchy induced by supervenience, and there is a horizontal ordering induced by causality at the physical level, so that mental and physical events are synchronized with each other (Fig. 5.1). This basic theoretical framework, often referred to as *minimal physicalism*, does make no commitment as to whether mental states are reducible to physical states, or whether there is a causal relation between mental states and between mental and physical states.

Apart from the supervenience relation, the physicalist also believes that the physical world is *causally closed*, that is, any physical event can only have another physical event as its cause. So it would seem impossible to have a physical event P as a cause for another physical event Q, and at the same time also having a mental event M causing the same physical event Q without running into troubles with *systematic overdetermination*. The physical event does all the "causal work", so there is nothing left to be done for the mental event. This dilemma has been precisely nailed down by Jaegwon Kim as the *problem of mental causation*: Causal efficacy of mental properties is inconsistent with the joint acceptance of the following four claims: (i) physical closure, (ii) causal exclusion, (iii) mind-body supervenience, and (iv) mental/physical property dualism – the view that mental properties are irreducible to physical properties (Kim 2005, 21).

Reductive physicalism avoids Kim's dilemma (or actually his pentalemma, involving the four claims plus a "zeroth" claim for the causal efficacy of mental states) by denying (iv), that is by holding that mental properties are nothing over and above physical properties. On the contrary, *epiphenomenalism*, which is a kind of non-reductive physicalism, holds that mental properties exist as standalone, non-reducible properties, but they would have no causal efficacy, so an epiphenomenalist denies the zeroth claim in Kim's pentalemma. Nevertheless, so the epiphenomalist would admit, the notion of mental states is useful as an epistemic conception to

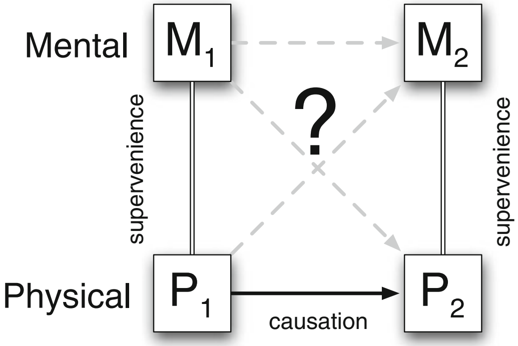


Fig. 5.1 Causal structure in minimal physicalism. The mental properties supervene on the physical properties, there is causality at the physical level, whereas causality at the mental level or between the levels is controversial

conveniently describe our actions at higher levels. Ontologically, however, mental states would be like a shadow or a symbol of the brain states:

> It seems to me that in men, as in brutes, there is no proof that any state of consciousness is the cause of change in the motion of the matter of the organism. If these positions are well based, it follows that our mental conditions are simply the *symbols in consciousness* of the changes which takes place automatically in the organism; and that, to take an extreme illustration, the feeling we call volition is not the cause of a voluntary act, but the *symbol of that state of the brain* which is the immediate cause of that act. (Huxley 1874, our emphases)

Also epiphenomenalism can be differentiated into a type and a token variant:

> According to the event- or token-epiphenomenalism defended by Huxley, concrete physical events are causes, but mental events cannot cause anything. According to the kind of property- or type-epiphenomenalism that threatens modern non-reductive physicalism, events are causes in virtue of their physical properties, but no event is a cause in virtue of its mental properties. (Walter 2007)

There is a great amount of resistance amongst philosophers against epiphenomenalism, and also Pettit, who is committed to a functionalist version of physicalism, seeks to avoid this rather unpopular kind: “We think, however, that there is an important error in the line of thought that suggests that functionalism makes content (and mental properties in general) causally irrelevant or epiphenomenal” (Jackson and Pettit 1990, 199).

In the next section, we introduce Pettit’s program model and the solution that it may provide to the problem of mental causation.

5.2 Pettit’s Program Model

Pettit proposes a model of causation that he calls the *program model,* which is set out to clarify the causal structure at different levels of description. An essential feature of the program model is its *hierarchical structure* with higher-level and lower-level properties that program for certain events, thereby realizing intra- and inter-level causality. The levels are connected with each other by a supervenience relation, so that a higher-level property supervenes on a lower-level property.

It appears to us that Pettit tacitly follows a conception of events that has been introduced by Jaegwon Kim (Kim 1976), and which defines an event as the instantiation of a certain property or set of properties at a particular time or during a particular period of time. Based on such conception of events, Pettit’s model would analyze the causal relation between two events A and X as the presence of at least one property of event A that *programs for* the production of event X. Any such A-property would then be denoted as *causally relevant* for the production of X. This conception of type causality is complemented by an analog conception for token causality. One particular instance, or token, of a causally relevant A-property would *program for* a particular occurrence, or token, of X, and this property-instance of A would be denoted as *causally efficacious* for that event-instance of type X. This idiosyncratic distinction between causal relevance and causal efficacy is given explicit import by Pettit: "Property-instances, and only property-instances, have causal effcacy; properties, and only properties, have causal relevance" (JD, 220) (Fig. 5.2).

Once that a property A has been identified as causally relevant for the production of a certain event of type X, it might be the case that there are *specifications a,b,c* of A some of which are also causally relevant for the production of X. Taken the other way round, A would represent a *generalization* of each of the properties *a,b,c*, and according to the program model, each of these may program for the same X-type event without getting in conflict with each other. The relation of specification (downwards) and generalization (upwards) induces a hierarchy on the properties that is taken as the basis for multilevel causation, eventually allowing for more powerful explanations:

> The methodological lesson is that we should eschew explanatory reductionism, which would favor going to finer and finer grains of information in seeking the explanation of any event. We should embrace what I call "explanatory ecumenism" or pluralism. (JD, 226)

5.2.1 The Lightning Example

Pettit provides the example of a bolt of lightning causing the electricity in a particular house to fail (JD, 218). Once it is justified to say that there was a bolt of lightning that caused the electricity to fail, and given the fact that this bolt of lightning happened to be of a certain magnitude *m*, then it is justified to also say that a bolt of lightning of magnitude *m* caused the electricity to fail. Hence, at the level of types we would have it that both the property of being a bolt of lightning and the property

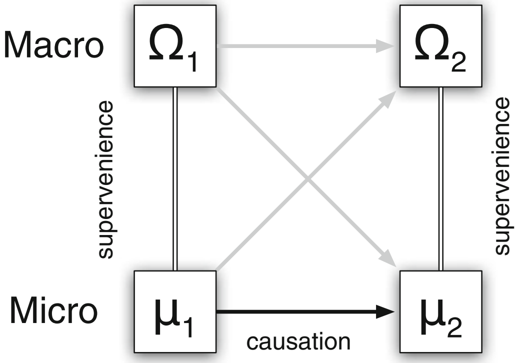


Fig. 5.2 Causal structure in Pettit's program model applied to a physical scenario involving two levels. Higher-level (*macro*) properties supervene on lower-level (*micro*) properties. Causality at the lower level is guaranteed by physicalism, whereas causality at higher levels and between the levels is derived from the program model

of being a bolt of lightning of magnitude *m* would be *causally relevant* for a failure of electricity in the house. At the token level, the instantiations of the higher-level and the lower-level property would both be *causally efficacious* for the electricity failure. According to the program model, there is no conflict between the different levels of causation; the bolt of lightning does not "steal away" the causal work from the bolt of lightning of magnitude *m* or vice versa. Both properties are *co-programming* the electricity failure in the house. Let us remark that in this example the higher-level properties would be *reducible* to the lower-level properties: a higher-level property is nothing *over and above* each of its specifications.

5.2.2 The Flask Example

The lightening example might appear slightly trivial as the higher-level and the lower-level properties are *co-instantiated*: A bolt of lightning of magnitude m is a *possible realization* of a bolt of lightning, and an instance of a bolt of lightning of magnitude *m* is also an instance of a bolt of lightning. Things get less trivial when Pettit claims that also in the case where two properties are *not* co-instantiated, they can still co-program for an event. For illustration he presents the example of a glass flask filled with boiling water (JD, 223). At some higher level, the macro level, the boiling of the water eventually causes the flask to crack. At the lowest level, the micro level, a particular molecule increases its vibration so strongly that it breaks the molecular bond in the flask. Which of these two causal stories is true? Pettit says: both are true, as the increasing of the vibration of the culpable molecule is a *component event* of the boiling of the water. The water's property of boiling *programs for* the cracking of the flask, and there is also a property of some component event of the boiling of the water, here the increasing of the vibration of a certain molecule, which *programs for* the cracking of the flask. Both property-instances are *causally efficacious* for the cracking of the flask. Thus, we do not only have a hierarchical structure of descriptive layers, as in the lightening example, but also of parts and wholes. Let us remark that also for this kind of scenario the higher-level properties would be *reducible* to the lower-level properties, as the former are *composed* of the latter: The boiling of the water is nothing over and above the vibration of its molecules. There might be problems with the intermixture of hierarchy concepts in the program model, as has been pointed out by Hüttemann and Papineau (2005), but we will not follow this route here, as we aim for something else. Let us buy it that if a certain event A causes another event X, then it is allowed to also say that there is a component event of A that causes X, and that both causal stories are true and not in conflict with each other.

5.2.3 A Solution to the Problem of Mental Causation?

The bottom line of the program model is that systematic overdetermination does not pose a problem for causation if the candidate events are hierarchically connected in a certain way. This sounds like a promising basis for a solution of the problem of mental causation. According to physicalism, there is a supervenience relation between the mental and the physical, implying a hierarchical structure:

> The supervenience reply to our problem, thus, is the observation that from the fact that the whole causal story can be told in neurophysiological terms, and that no functional property is any neurophysiological property, it does not follow that the functional properties do not appear in the story. They appear in the story by supervening on the neurophysiological properties […]. (Jackson and Pettit 1990, 201)

Pettit explicitly suggests that the program model can and should be applied to the case of mental causation:

> It is important to see that the program model may be extended to more complex cases like this. I think that the model helps to make sense of how we can have mental or intentional causation at the same time as physical causation, and social causation at the same time as intentional causation, where intentional properties are higher in level than physical ones, and social properties are higher in level than intentional ones. (JD, 223)

Given that we interpret some property of a mental event of type M as being higher-level with respect to some property of a physical event of type P due to the fact that the respective M-property supervenes on the respective P-property, and given that this P-property is causally relevant for the production of another physical event of type Q, then we can invoke the program model to justify the claim that the respective M-property is *also* causally relevant for the production of Q. At the token level we would have an analog conclusion, hence we would have a justification of both type-causality and token-causality in the case of mental and physical events by identifying the systematic overdetermination in this case as unproblematic: it's not a bug, it's a feature!

In this section we presented Pettit's program model and how it may be used to solve the problem of mental causation. Pettit seems to use Kim's conception of events, and he holds that only property-instances are causal efficacious. Pettit provides two examples to illustrate his conception: one of a bolt of lightning, where the properties are co-instantiated, and another one of a flask filled with boiling water, where the properties are not co-instantiated. We point out that in both scenarios the higher level properties are reducible to the lower level properties in the sense that the former are nothing over and above the latter.

5.3 Objection

Our main objection to Pettit's solution to the problem of mental causation is the following. His program model runs danger of clarifying only the *epistemic* aspects of causation, leaving their *ontological* status unexplained. We have no problem accepting that it makes sense to allow for the truth of different causal stories at different levels of description. We are even happy with Pettit's approach as it explains why our intuitive conception of causality makes sense even though we are most of the time *not* referring to the lowest level of physical description, that is, to particles, fields and forces. We take it that the concept of "causal work" which can be somehow used up by lower-level events, does not really make sense at the descriptive level, at least not with respect to the intuitive meaning of causality. One can still be true with saying that the kicked football broke the window, with or without knowing that there is *actually* a heap of particles which form a football and which interact with other particles forming a window in such a way that the molecular bonds of some of the window particles are broken. One does not need to know all this, and even if one knew all this, the simple causal story about a football breaking a window would still remain true. The process that is going on may be described at different levels as a causation between certain events. But, and here is our concern, to *describe* something does not necessarily mean to have captured what is going on *ontologically*. There is no consensus about what mental properties, states and events really *are* and how they are related to physical properties, states and events. Yet, it is not even clear if the notion of causality is a reasonable concept beyond the epistemic domain (see Russell 1912; Rheinwald 2012, for a criticism of causality as an extensional notion). The program model, if taken seriously as a metaphysical framework, would somewhat magically bridge the gap between merely convenient descriptions and their underlying ontology.

5.3.1 The Pong Example

To strengthen our concern described above, we are going to illustrate by a counterexample that a supervenience relation alone does not suffice to justify the application of the program model, at least not with a satisfying result. In this counterexample it will be intuitively clear that the supervening phenomena are *in fact* mere epiphenomena without causal efficacy (Fig. 5.3).

Consider Jones playing the famous old-fashioned computer game Pong. It would seem to Jones that the contact of the "ball" with the "racket", displayed on the monitor as respectively a small square and a lengthy rectangle, somehow causes the ball to be reflected to the opposite side of the screen. In fact, Jones would act just *as if* this little causal story were true. He will move the "racket", by using the mouse or the computer keyboard, so that it comes into contact with the "ball" and be reflected to the opposite side of the screen. Now let us tentatively apply Pettit's program

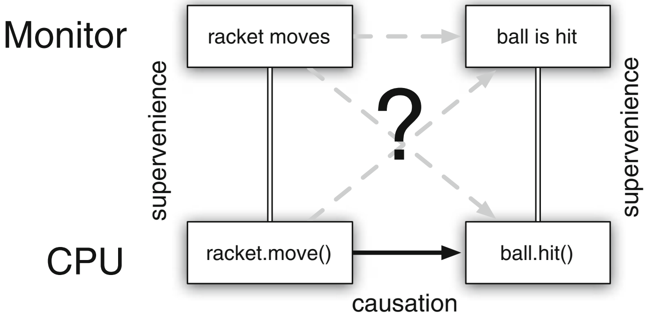


Fig. 5.3 The Pong example. The graphical states of the computer monitor supervene on the electronic states of the computer CPU. The causality at the lower level is intuitively clear. Can the program model be applied to account for causality also at the level of the graphical display and between the levels?

model. The position and velocity of the "ball" and the "racket" on the monitor are physical properties that supervene on the physical properties of the CPU. According to the program model, some of these physical properties of the CPU would *program for* an electronic event of a certain type – let us denote it here as "ball.hit()" in resemblance to a computer language – which is taking place in the CPU and which gives rise to a corresponding graphical event of a certain type – let us denote it as "ball hit" – on the computer screen. At the same time, the movement of the "racket" on the monitor would *also* program for the same electronic event of type "ball.hit()" *and* it would program for the corresponding graphical event of type "ball hit", since the graphical event is instantiated together with the corresponding electronic event. With respect to what has caused the electronic event "ball.hit()", there are two causal stories to be told: one being the electronic story taking place in the processing hardware of the computer, and the other being the graphical story taking place on the monitor. According to the program model, both causal stories would be true. Yet it would seem intuitively that only one of these stories is true from an ontological point of view, and it is the electronic story and not the graphical story. Still, from an epistemic point of view it might be useful and convenient to attribute causal efficacy to virtual events taking place on a computer monitor, and this is in fact what all of us do all the time when we interact with computers. We would simply say "when the racket moves like this it causes the ball to be hit", but we would not seriously hold that those pixels on the screen do any *real* causal work to the other pixels or even to the electronic state of the CPU. Rather, the visual occurrences on the computer monitor are mere *epiphenomena* without causal efficacy; they only *indicate* what is going on at the electronic level.

5.3.2 Implications

The Pong example illustrates that if a supervenience relation between different properties were taken to be sufficient for treating them as higher-level and lower-level properties in the sense of the program model, then we may happen to ascribe the higher-level properties causal relevance although they are *in fact* mere epiphenomena. As a consequence, Pettit's program model, when applied to mental causation, would be compatible with epiphenomenalism.

Another possibility is to deny that the supervenience relation between different properties suffices for treating them as higher-level and lower-level properties in the sense of the program model. Indeed, Pettit's examples only consider cases where the lower-level properties are either specifications of the higher-level properties (as in the lightening example), or where the lower-level properties are component properties of the higher-level properties (as in the glass flask example). Epiphenomena, however, are neither generalizations of physical entities nor do they stand in a part-whole relation with them. Also in our Pong example, the epiphenomenal graphical representations on the computer's monitor are neither generalizations of the electronic states of the computer's CPU, nor do they stand in a part-whole relationship with the latter. Still, the graphical representations supervene on the CPU states, just as mental states supervene on physical states. So, a defender of the program model might argue that our counterexample misses the point, and that there is more needed than a supervenience relation to constitute a hierarchy in the sense of the program model. However, since a supervenience relation between mental and physical states is all that is assumed in minimal physicalism, the program model could then simply not be applied to mental causation, at least not on grounds of minimal physicalism alone.

A third possibility is to hold that mental properties are *in fact* either generalizations of physical properties or that they are in some way composed of the latter. Doing so, one would obtain a hierarchical structure suitable for the application of the program model, but then the resulting theory would be a version of reductive physicalism, as mental events would then be reducible to physical events.

5.4 Conclusion

Altogether, the options for dealing with our objection raised in the previous section are the following:

1. The program model is only applicable to certain kinds of causal scenarios, but it is not applicable to mental causation.
2. The program model is applicable to all kinds of causal scenarios including mental causation, but it implies a version of reductive physicalism.

3. The program model is applicable to all kinds of causal scenarios including mental causation, but it remains merely descriptive and is compatible with epiphenomenalism.

If one is not satisfied with either of these options, and one still wants to apply the program model to mental causation, then, so is our point, it takes more or stronger premises than are provided by minimal physicalism to be safe from epiphenomenalism, and it takes less or weaker premises than are implicit in Pettit's examples to be save from reductive physicalism. Which, then, are these premises?

References

Hüttemann, A., and D. Papineau. 2005. Physicalism decomposed. *Analysis* 65(285): 33–39.

Huxley, T.H. 1874. On the hypothesis that animals are automata, and its history. *The Fortnightly Review* 16: 550–580.

Jackson, F. 1998. *From metaphysics to ethics: A defence of conceptual analysis*. Oxford: Oxford University Press.

Jackson, F., and P. Pettit. 1990. Causation in the philosophy of mind. *Philosophy and Phenomenological Research* 50: 195–214.

Kim, J. 1976. Events as property exemplifications. In *Action theory*, ed. D. Walton and M. Brand, 159–177. Dordrecht: Reidel.

Kim, J. 2005. *Physicalism, or something near enough*. Princeton: Princeton University Press.

Pettit, P. 1995. Causality at higher levels. In *Causal cognition: A multidisciplinary debate*, Fyssen Foundation, ed. D. Sperber, D. Premack, and A. Premack, 399–421. Oxford: Clarendon.

Rheinwald, R. 2012 [1994]. Causation and intensionality: A problem for naturalism. In *Rosemarie Rheinwald: Logik, Kausalität, Freiheit – Ausgewählte Aufsätze/Logic, Causality, Freedom – Selected Papers*, ed. J.G. Michel and O.R. Scholz. Paderborn: Mentis.

Russell, B. 1912. On the notion of cause. *Proceedings of the Aristotelian Society* 13: 1–26.

Walter, S. 2007. *Epiphenomenalism*. Internet Encyclopedia of Philosophy. http://www.iep.utm.edu/epipheno/. Accessed 28 Aug 2015.

Chapter 6
Notes on Pettit's Concept of Orthonomy

Alexa Nossek and Julia Belz

6.1 Introduction

Freedom plays an important role in Philip Pettit's work. He tries to establish a rich concept of freedom which contains psychological as well as political aspects. One part of his work is a new interpretation of the notion *autonomy*[1] which he calls *orthonomy*. This concept is established by Pettit and Michael Smith in three different essays discussing *orthonomy* in several contexts. The first time the notion *orthonomy* appears in *Backgrounding Desire* as an implication of the so called strict background view of desire. *Orthonomy* basically is a substantive concept of autonomy. Contrary to a content-neutral, purely procedural concept it requires that someone to whom autonomy is ascribed exercises certain values. In the roughest definition being *orthonomous* means being guided by what is right (JD, 239).

We think that there are some problems and open questions connected to the conception of *orthonomy*. We question the thesis that *orthonomy* is an implication of the strict background view of desire (Sect. 6.4.1), speak of the problematic connection of desires and values (Sect. 6.4.2) and ask whether values in Pettit's definition can be objective in any way (Sect. 6.4.3). We will present this after a rough depiction of the strict background view of desire (Sect. 6.2) and a short description of the idea of *orthonomy* and its social dimension called *conversability* (Sect. 6.3).[2]

[1] If not explicitly said otherwise, we understand *freedom* as a broad notion and *autonomy* as a certain feature of freedom, which – of course – needs to be defined. Pettit sometimes speaks of autonomy, sometimes of free will (or free will and free thought). He seems to use both expressions synonymously (JD, 238 f.).

[2] Because of the limited space we are not able to give a complete picture of Pettit's account of *orthonomy*. We do not refer to the metaphysical (Pettit and Smith 1996, 444; JD, 240) and metaethical assumptions (Pettit and Smith 1996, 442 f.). We analyze the concept of value just insofar as

A. Nossek (✉) • J. Belz
Philosophisches Seminar, Westfälische Wilhelms-Universität, Münster, Germany
e-mail: alexanossek@uni-muenster.de; juliabelz@gmx.de

S. Derpmann, D.P. Schweikard (eds.), *Philip Pettit: Five Themes from his Work*,
Münster Lectures in Philosophy 1, DOI 10.1007/978-3-319-26103-4_6

6.2 The Strict Background View of Desire

6.2.1 Description

According to the strict background view of desire, desires appear always in the background of a person's deliberative decision-making without simultaneously being in the foreground of her decision-making (Pettit and Smith 1990, 566).[3] Pettit and Smith define the difference between a desire being present in the background and being present in the foreground as follows:

> More generally, a desire is present in the background of an agent's decision if and only if it is part of the motivating reason for it: the rationalizing set of beliefs and desires which produce the decision. A desire is present in the foreground of the decision if and only if the agent believed he had that desire and was moved by the belief that a justifying reason for the decision was that the option chosen promised to satisfy that desire. A desire may be in the background without in this sense figuring in the foreground. And equally a desire may be in the foreground without being in the background. (Pettit and Smith 1990, 568)

To put it short, one can say that usually each decision includes deliberation and involves a desire which motivates the person's decision to act in a certain way. Only if the choosing person also has the belief that she has the desire and also meets the other requirements mentioned in the quote above, the desire figures in the foreground additionally (Pettit and Smith 1990, 568).[4]

To understand this account properly, two points have to be made clear. First, the notion of desire always refers to a type, not a token. Second, the distinction between the background and the foreground is strictly functional, i.e. it has nothing to do with the difference between the conscious and the unconscious (Pettit and Smith 1990, 568).[5]

we need it for our topic here, neglecting the interesting connection to virtues and the Aristotelian theory of the middle way (Pettit and Smith 1993, 76 f.).

[3] For Pettit's understanding of deliberation, Pettit 2010. In his article he presents his thesis that, firstly, each decision is an act and that, secondly, there is no decision without deliberation (Pettit 2010, 255).

[4] The case indicated above that there is a desire in the foreground without being in the background refers to the possibility that one could be mistaken in thinking that the desire truly motivates the decision.

[5] Pettit's and Smith's theory is called the *strict* background view of desire because of their thesis that the desire always is in the background but not always is in the foreground. The counterpart which the authors do not accept would be the so called loose background view of desire according to which a desire is always present in the background and in the foreground simultaneously (Pettit and Smith 1990, 572 f.).

6.2.2 The Implication on Autonomy

As mentioned above, the strict background view of desire is supposed to require a new interpretation of autonomy, i.e. the alleged truth of this theory of desire implies the model of *orthonomy*. As a reason for this thesis the Pettit and Smith refer to the point that if the desire principally figures in the background, the scope of a desire-based theory of autonomy is reduced (Pettit and Smith 1990, 585 f.).[6]

In some remarks[7] Pettit and Smith indicate that their philosophical opponents are to be found in the existentialistic position of Jean-Paul Sartre and the *higher-order-theory* of Harry G. Frankfurt which is a desire-based, content-neutral conception of autonomy (Frankfurt 1998). They emphatically refuse the idea of autonomy consisting in the identification with a desire by endorsing it in a mental state of a higher level.[8] The reasons for the rejection of hierarchical analyses of autonomy are the classical criticisms discussed in the debate about autonomy. The authors mention the threat of an indefinite regress of higher orders and the danger of expecting too much from persons, because one can understand the *higher-order-theory* in such a way, that every single desire has to be examined in order to be proven autonomous (Pettit and Smith 1990, 586). Moreover, Pettit and Smith object that the second-order desires or volitions might be generated by some kind of conditioning during the childhood (Pettit and Smith 1996, 443). In *A Theory of Freedom,* Pettit does not completely reject the idea of higher-order volitions, but claims that they are only plausible in an interpretation related to his theory of freedom as *discursive control*. It is important to understand that those mental states of a higher level obviously are not necessary for autonomy (TF, 90).

According to Pettit and Smith, the crucial question regarding autonomy is: What does it mean for an agent to be master of his/her desires, "not letting them undermine his values?" (Pettit and Smith 1990, 586) They conclude that if desires were to figure always in the foreground, then each desire had to be analyzed in order to prove it to be autonomous: "Autonomy is going to be at risk in every act of decision then." (Ibid.)

Accepting the strict background view of desire autonomy is threatened only in a few cases. Pettit and Smith refer to akrasia (weakness of will), compulsion and caprice. These are the standard cases which occur within the debate about autonomy. To put it more clearly, the authors claim that beyond such pathologies there are no situations in which the autonomy of a person can or should be questioned (Pettit and Smith 1990, 587 f.).

[6] Note that a decision-based theory of autonomy might be completely unaffected by the so called background-foreground-question. e.g. Taylor 2009.

[7] Pettit and Smith (1990, 586f, 1993, 76, 1996, 442 f).

[8] "And equally we see things very differently from someone like Harry Frankfurt, who requires any operative desires, or at least any operative ground-level desires, to be desires that are endorsed a level up: desires that the agent desire to act on[…]." (Pettit and Smith 1993, 76) TF, 53–57.

6.2.3 Values in Pettit's Thought

As shown above, autonomy is meant to consist in a harmony between desires and values. Pettit and Smith continue with a definition of values: "Under the strict background view your values are naturally taken as the things you judge desirable, the properties you focus on in the course of deliberation." (Pettit and Smith 1990, 587)

This means that values are judgments and results of deliberation. Values can be undermined by desires, in Pettit's and Smith's words, "[…] that can only mean that some of your desires are not properly responsive to the values you hold: the properties you judge desirable." (Ibid.) It is puzzling that some sentences later the authors describe values as "[…] the properties which the agent countenances as valuable – the properties which he desires […]." (Ibid., 588) At this point there seems to be no difference between the object of a desire and something that is judged as desirable anymore.

The concept of *value* seems to be indistinct. Sometimes the authors put it in relation to virtues (Pettit and Smith 1993, 76 f.), sometimes it is something that a person wishes. Thus, it is unclear to what extent values are to be understood as objective or subjective.

Pettit and Smith give an example for an autonomy-threatening, pathological desire which undermines the values of an agent. Someone suffers from a compulsion to clean his room. The desire to make up his room outweighs all the person's other values and prevents the agent from getting to work punctually. According to the authors the desire to clean becomes the person's master making him a slave, "[…] my desires take charge of my values, if indeed I can be ascribed any values." (Pettit and Smith 1990, 587)

It is this interpretation of autonomy which connects the desires being in the background with values that the authors call *orthonomy*.

6.3 Orthonomy

6.3.1 The Definition of Orthonomy

In *Backgrounding Desire* Pettit defines *orthonomy* as follows:

> Under this interpretation [this refers to the strict background view of desire], a better name for the virtue might be 'orthonomy' rather than 'autonomy.' It consists in forming your desires according to the right sort of principles rather than the wrong. (Pettit and Smith 1990, 588)[9]

The primary formulation of *orthonomy* defines it as correspondence of one's desires and the right principles.

[9] It is unclear what Pettit and Smith mean by referring to a *sort* of principle.

In *Joining the Dots,* Pettit presents his thesis that freedom is connected with a capacity which he calls *orthonomy*. This capacity is clearly linked to a certain understanding of rationality.[10] Pettit refers to the so called demands of reason which one has to realize and which should guide one's acting. He explains:

> The discipline of reason is the discipline of what is right or *orthos* according to the reasons available to the agent; what 'right' means, in this usage, is that which has the support of relevant reasons So why not hold that free will should be identified with what Michael Smith and I have called 'orthonomy' [...]; that is, with the capacity to be guided by the orthos or 'right', according to available reasons. (JD, 238, emphasis in the original text)

It is important to bear in mind first that autonomy (in the interpretation of *orthonomy*) is defined as a capacity and second that *right* means supported by the relevant reasons.

As *right* is defined as supported by the relevant reasons, its concrete meaning is vague. At first it is unclear whether right means good in a moral sense or simply rational. But there is a passage which suggests that *right* actually includes a moral aspect. Pettit and Smith quote Gary Watson's example of a man who is angry because of his defeat in a game of squash and who therefore wants to smash his opponent's face. Pettit argues that since the player is unable to find any justification for that action, he has to refrain from it.[11]

Freedom consists in forming one's beliefs on the basis of a reason and act for reasons in accordance with the rules of rationality. What is important in Pettit's view is not the self-determination of an action or a belief, but "'Orthonomy' means the rule of the right rather than the rule of what is not right." (JD, 239) The antonym of both notions (orthonomy and autonomy) is heteronomy, which means the "rule of the non-right" (ibid.). In *Backgrounding Desire* the notion of heteronomy is explained a slightly different way. A heteronomous person acts at different times according to different values although there is no relevant change in the state of affairs (Pettit and Smith 1990, 588).

Pettit and Smith formulate requirements for *orthonomy*:

> We can be a little bit more precise on what orthonomy in this sense requires. It requires that the properties which the agent countenances as valuable – the properties which he desires – figure consistently in the determination of which option he comes to desire. (Pettit and Smith 1990, 588)

If one tries to reformulate this, one can say that: to be *orthonomous* a person has to decide in accordance to her values consistently, i.e. without being unstable or moody.

[10] Pettit and Smith 1993 and JD, 243 ff.

[11] Pettit and Smith 1996, 439 f. The authors mention the example in the context of freedom in desire.

In each case in which a person has to choose between several options, she can succeed or fail in practicing *orthonomy*. It is important to see that the authors characterize *orthonomy* as an ideal.[12]

6.3.2 Conversability

In *Freedom in Belief and Desire,* Pettit and Smith introduce the notion of the *conversational stance* and continue to develop their concept with the aid of this notion. This notion of the *conversational stance* shows that through talking to each other, persons develop and survey their believes (Pettit and Smith 1996, 440–442).[13]

This points to a social dimension of *orthonomy*. In *Joining the Dots* Pettit discusses this dimension under the notion of *conversability. Conversability* consists in the capacity to hold a conversation. Pettit says that a person must necessarily assume that her partner is able to reflect on his thinking and acting, to recognize reasons and, by doing this, act according to them and control himself. This shows that to converse with someone means that one recognizes him as a subject worth of participating in collaborative, rational thinking. Pettit calls this kind of thinking *co-reasoning*. Thereby co-reasoning with someone means nothing else than to regard him as *orthonomous* (JD, 274 f.).

Conversability – and this leads to the assumption that it also counts for *orthonomy* – exists as a self-characterization and also as a characterization in the social world (JD, 277). This means you can co-reason with yourself, as well as with someone else.

At this point, Pettit says again that to regard someone as *orthonomous* does not mean that he always acts *orthonomous*. It just means that he has this capacity. *Conversability*, just as *orthonomy,* is an ideal, which often remains unreached.

6.4 Criticism and Open Questions

6.4.1 Does the Strict Background View of Desire Really Imply Orthonomy?

We do not understand why the strict background view of desire should imply *Orthonomy* as the correct model of autonomy. The step from desires functionally being in the background to values as something that grants autonomy is not evident.

[12] Pettit and Smith 1993, 77: "As an ideal of pure practical reason, it proclaims its incompleteness on its face; it does not suggest, as the ideal of autonomy has sometimes done, that it represents the be-all and the end-all of morality. And that, surely, is to its credit." also JD, 240.

[13] In this context, the *conversational stance* and *conversability* also refer to the concept of *discursive control* (TF, 65 ff.).

Pettit and Smith have to explain their thesis and argue for it. Nothing has been said to make the acceptance of a substantive view of autonomy plausible. Desires figuring in the background of a process of decision-making have nothing to do with doing the right or doing something good.

We think that a procedural, content-neutral account of autonomy as well as many other conceptions could be fully compatible with the strict background view of desire. As an example one can take James Stacey Taylor's account of autonomy, who concentrates on decisions instead of desires. According to this view, a person has to be satisfied with the procedure of her decision-making to count as autonomous (Taylor 2009, 8).[14] Why not regard, for instance, Taylor's account as the one which is the correct interpretation of autonomy under the constraints of the strict background view of desire? The assumption that acting according to certain values is necessary to be autonomous is not evident. That is why we think that Pettit has to provide a stronger argument for this assumption.

6.4.2 Is Pettit's Connection of Desires and Values Correct?

Remember the example of a person who suffers from a compulsion to keep her room tidy. We doubt that the described case really points at a conflict between a desire and a value – in Pettit's definition of value. Pettit and Smith admit that tidiness can be seen as desirable and speak of the value that is ascribed to tidiness (Pettit and Smith 1990, 587). We think that the situation can be seen as a conflict between values – the value of tidiness and, for instance, the value of getting to work timely – on one level of description and as a conflict between desires on a different level. In the concrete situation the person suffers from a conflict between two desires. She wants to leave the room (in order to get to work) and she wants to stay and clean. Both actions must be described as desirable for the agent. Because how could it even be possible to desire something that is not desirable at all for the desiring agent?

To enlighten our thesis that it is impossible to desire something which is not desirable for the desiring agent, consider the example of a diabetic, who wants to eat some chocolate. The relevant reasons available to the patient – in this case the profound, medical advice – clearly do not support the action of eating the candy. Therefore, the agent judges it to be undesirable. Moreover, he wants the desire to

[14] It is impossible to give an adequate description of Taylor's view in this paper. Taylor claims to present an analysis of autonomy (Taylor 2009, 2) which consists in "[…] a decision-based, historical, externalist, and political account of personal autonomy […]" (Taylor 2009, 17). Moreover, his theory is "value-neutral" (Taylor 2009, 17), i.e. it is not a substantive but a content-neutral conception of autonomy. The only exception Taylor makes is his thesis that in order to be autonomous one has to appreciate a certain degree of critical reflection and one's own autonomy. Only in this sense Taylor's account can be called minimally substantive (Taylor 2009, 70). Note that Taylor's use of the notion *decision* is compatible with Pettit's definition (Pettit 2010, 253 ff.). We do not want to follow Taylor's view in any respect. It just serves as an example.

refrain from eating chocolate to be effective (or operative in Pettit's terminology). But nevertheless he still desires eating the candy. And this very desire means that on another level he judges the consumption to be desirable. The reason for this different judgment is the pleasure of enjoying the taste and texture of the chocolate.

A concept of autonomy for real persons has to accurately picture what goes on in human agents, and it has to regard the different levels, the hierarchy of mental states within an agent. In the person's inner hierarchy of values one value can be regarded as more important than another. In the portrayed case the value to go to work on time is ascribed more weight by the agent. This deliberation shows that Pettit and Smith are mistaken when they conclude that maybe such an agent cannot be ascribed any values anymore.

The matter is far more complex than Pettit and Smith put it. It seems to us that Frankfurt's *higher-order-theory* expresses what the authors really mean in the depicted context. The person wants to leave the room and she wants to stay. Furthermore the person wants to want to leave the room, i.e. she suffers from a conflict of first-level desires but has formed the second-order volition to leave the room. The authors' explicit rejection of a theory of higher levels leads to an inadequate description of what goes on in an agent who suffers from an inner conflict. Clarification and a more adequate view of desires and values are needed in order to improve the theory of *orthonomy* and to make it suitable for real persons.

6.4.3 Are Values Objective in Any Way?

There seems to be another problem regarding the role of values in Pettit's account of *orthonomy. Value* is defined as something which a person judges as desirable (Pettit and Smith 1990, 587; see also above). The judging person not only thinks that this is desirable for herself, but also regards the value as counting for other people. Nevertheless, what I consider as desirable is in the first instance only desirable for me. The introduction of the conversational stance and the conception of *conversability* seem to help. People form desires and values by talking to each other. In a conversation they ask for reasons and give them to each other. But in spite of the tendency to globalization, this procedure of building values only works in an individual community. There are different values in, say, Germany, the USA and China. Moreover, within one country there are kinds of different societies following different values. In Germany, for example, the mass media often worry about so called *Parallelgesellschaften* build up by orthodox Muslims who want to obey only their own religious rules breaking with the values of the majority of the German society. Pettit makes clear that someone is only *orthonomous* if he is rational.[15] But many agents and communities with totally different values will pass the threshold of rationality.

[15] Remember in this context that *right* is defined as supported by the relevant reasons (JD, 238).

Another question occurs that is related to Pettit's theory of *orthonomy*. Does the explained consideration hint at the fact that in view of the apparent plurality of values there are countless versions of *the right* different individuals and communities act according to? In our opinion, Pettit should clarify his theory of *orthonomy* in order to deal with the plurality of values. A strictly subjective view of values would be problematic. If every person has different values that do not count for other people, then it remains unclear what *right* means. The whole discussion about the conversational stance would not make any sense, because there would be no point in co-reasoning when each person has different values and different definitions of what is right. In a broader dimension the same is true for different communities and societies. There has to be a more objective view of values in order to have a shared definition of what is right and of what is valuable.

References

Frankfurt, Harry G. 1998. Freedom of the will and the concept of a person [1971]. In *The importance of what we care about. Philosophical essays*, 11–25. Cambridge: Cambridge University Press.

Pettit, Philip. 2010. Deliberation and decision. In *A companion to the philosophy of action*, ed. T. O'Connor and C. Sandis, 252–258. Oxford: Blackwell Companions to Philosophy.

Pettit, Philip, and Michael Smith. 1990. Backgrounding desire. *The Philosophical Review* 99(4): 565–592.

Pettit, Philip, and Michael Smith. 1993. Practical unreason. *Mind New Series* 102(405): 53–79.

Pettit, Philip, and Michael Smith. 1996. Freedom in belief and desire. *The Journal of Philosophy* 93(9): 429–449.

Taylor, James Stacey. 2009. *Practical autonomy and bioethics*, Routledge annals of bioethics, eds. Mark J. Cherry and Anna Smith Iltis. New York: Routledge (First published 2009 by Routledge).

Chapter 7
Two Problems of Value-Monistic Consequentialism in Philip Pettit's Theory of Criminal Justice

Tim Grafe, Tobias Hachmann, and Michael Sabuga

7.1 Introduction

In their book *Not Just Deserts. A Republican Theory of Criminal Justice* (1990) John Braithwaite and Philip Pettit formulate a theory of criminal justice based on an approach that is consequentialist and republican. This course is followed by Pettit in his later works *Republicanism* and *Republican Theory and Criminal Punishment* (both 1997). In this essay we want to illustrate the main aspects of this consequentialist theory of criminal justice and point at two problems this theory has to face. These problems are (i) its difficulty to take rights seriously and (ii) its vagueness confronting the user with difficulties in application. We will concentrate on *Not Just Deserts*, the most elaborated work on criminal justice under participation of Philip Pettit. But we will also give cross references to other works if necessary.

7.2 A Comprehensive Theory: Value Monism

Braithwaite and Pettit defend a "comprehensive" theory of criminal law. They do not want to give separate criteria for the justification of punishment exclusively, but rather try to find criteria for evaluating the criminal system, which consists of different "sub-systems" (besides punishment, e. g. of the allocation of resources among different systems parts or questions of surveillance and investigation, among others), as a whole (NJD, 12–27)[1]: "By 'comprehensive' we mean a theory which will give an integrated account of what ought to be done by the legislature, the judiciary, and the

[1] This is one reason for refuting retributivism as a theory of punishment. Retributivism holds that an offender has to be punished because he "deserves it" (NJD, 2). Punishment is determined solely

T. Grafe (✉) • T. Hachmann • M. Sabuga
Philosophisches Seminar, Westfälische Wilhelms-Universität, Münster, Germany
e-mail: tim.grafe@gmx.de; Tobias-Hachmann@gmx.de; msabuga@web.de

S. Derpmann, D.P. Schweikard (eds.), *Philip Pettit: Five Themes from his Work*,
Münster Lectures in Philosophy 1, DOI 10.1007/978-3-319-26103-4_7

executive in regard to the key policy questions raised by the criminal justice system." (NJD, 12). Such a comprehensive theory gives a set of answers to those questions that is (i) complete, (ii) coherent and (iii) systemic; i. e. it (i) gives answers to all of these questions, (ii) those answers are consistent and (iii) "derived in a process of continuing comparison and adjustment" (NJD, 15–16). The reason for this understanding is the idea that a separate evaluation of the 'sub-systems' can lead to inefficiency, or even counterproductivity (NJD, 7, 8, 17), as almost every particular measure can affect the functions of many 'sub-systems'. Sometimes a desired result can be achieved more effectively with the tools of other 'sub-systems'; in other cases measures on one area might simply contradict those on another one.

Using the method of "comparison and adjustment" of the various answers, different systems of criminal justice could be established. According to the authors, a normative theory of criminal law should not "consist simply in an enumeration and ordering of the different possible systems of criminal justice", but provide "a criterion by which such a ranking can be developed" (NJD, 25). What is sought for, therefore, is a supreme criterion for evaluating different possible systems of criminal justice. We shall call such a theory 'value-monistic'.

7.3 Consequentialism: Promoting, Not Honoring

Subsequently, Braithwaite and Pettit discuss the nature of the sought-after criterion. They distinguish primarily between two different kinds of criteria: Consequentialist and deontological ones. Both can appreciate the same value but they differ in how the value is used: Consequentialist criteria define "targets", deontological ones define "constraints" (NJD, 27). A 'constraint' is a value which must be "exemplified" – respectively "honored" (CP, 126) – by an actor's actions (NJD, 27). Actions must not violate this value. Rights, for example, are just the other side of constraints. If one person has got a right to the treatment X, other persons have the constraint to (i. e. have to) treat the right-bearer in the X-way, whatever else is the case.

A 'target', on the contrary, is a value that must be "promoted" (CP, 126). The value has to be increased in the consequences of an action. Therefore – here lies an important difference to deontology – an agent may even violate the value through his action in a particular instance as long as the value is promoted in the consequences in such a way that it is maximized overall (NJD, 27).

Braithwaite and Pettit argue in favor of a consequentialist criterion.[2] According to them, a deontological criterion is not sufficient with a view to formulate a comprehensive theory. It is only designed to consider particular questions and does not take into account the consequences that occur because of honoring the constraint (NJD, 30–31). Trying to consider all the areas of the system will therefore "inevitably

"in proportion to their desert", which mainly means "harmfulness" and "blameworthiness" (NJD, 4). (Supposed) social consequences of the punishment are not considered.

[2] The later theory follows a consequentialist criterion, too (R, 81).

bring a target into play", i. e. a further kind of criterion; furthermore, two kinds of criteria might be conflicting and one would have to decide when a certain kind of criterion should be applied and for what reason (NJD, 36–38). To avoid these problems, the authors decided to start directly with the consequentialist approach which, in their view, does not need further values of some other sort. Nevertheless, Braithwaite and Pettit admit that even a theory with an ultimate consequentialist target "must make a place for constraints at some less fundamental level" in order to be credible (NJD, 32). We shall come back to this point soon.

7.4 Three Requirements of a Monist Criterion

So what does the supreme aim for the criminal justice system have to look like? The authors formulate three "desiderata" that this aim has to exemplify. It should be "uncontroversial", "stabilizing" and "satiable" (NJD, 42–45). 'Uncontroversial' means that the target must be widely recognized among people within "Western-style democracies"; a 'target' is 'stabilizing' if it provides a "stable allocation" of 'uncontroversial' rights, as, for example, the right of the innocent not to be punished (NJD, 42–44). An allocation is stable, if it motivates actors in the criminal justice system to take these rights seriously (NJD, 72). Finally, a 'target' is 'satiable' if it provides not only a stable allocation of individual rights, but also does not violate general legal principles that are uncontroversial in western democratic societies, like the prohibition of an excessive surveillance of public places (NJD, 42–45).

With preventionism[3] and utilitarianism,[4] the authors refute two well-known consequentialist and comprehensive criteria for the criminal justice system. Both these theories do not exemplify the second and third desiderata, among others. Preventionism is rejected for being "outrageously destabilizing and insatiable" (NJD, 45), and utilitarianism can lead to massive intrusion into subjective rights as well as violate uncontroversial legal principles, too (NJD, 52–53). Furthermore, utilitarianism most likely does not satisfy the first desideratum, "because it is so vaguely defined" (NJD, 52).

[3] The supreme target of a preventionist theory is the prevention of crime, respectively the prevention of harm caused by crime (NJD, 45, 47).

[4] The supreme target of a utilitarian theory of criminal justice is the "maximization of the happiness of those affected by the criminal justice system" (NJD, 52).

7.5 The Promotion of Dominion: A Monist Criterion for the Criminal Justice System

The supreme target that Braithwaite and Pettit formulate is the promotion of republican freedom, which is called "dominion" (NJD, 9, 54, 87). 'Dominion' is "an individual-centered value", i.e. freedom of individual persons rather than that of collective entities (NJD, 55). 'Promotion' means that the *overall* 'dominion' has to be maximized. These considerations demonstrate two things: first, that the reduction of dominion of singular persons can be legitimate[5]; and second, which is more general that, while originally being introduced as a criterion for evaluating different sets of theories of criminal justice, 'dominion' serves as a criterion for evaluating particular measures, too.

How is dominion exactly defined? A person has "full dominion […] if and only if

1. she enjoys no less a prospect of liberty than is available to other citizens.
2. it is common knowledge among citizens that this condition obtains […].
3. she enjoys no less a prospect of liberty than the best that is compatible with the same prospect for all citizens." (NJD, 64–65)[6]

'Dominion' is a form of republican *negative* freedom, which consists in being "exempt from the constraints imposed by the intentional or at least blameworthy actions of others in choosing […] options which the normal agent is capable of realizing in normal conditions without the special collaboration of colleagues or circumstances" (NJD, 61). The republican specification of negative liberty consists in modifying the last part of this definition: freedom is not made of all the options you

[5] This holds explicitly for offenders to be convicted (NJD, 78). In his later work, however, Pettit introduces the distinction between factors *conditioning* and *compromising* freedom. Legitimate law does not compromise (violate) freedom, but solely conditions it, like natural obstacles do (Pettit 1997, 62–63). According to this view, punishment that is not arbitrary (see footnote below) does therefore not compromise the offender's freedom; the options made impossible by legitimate punishment were rather never guaranteed, they did never belong to freedom, as one could say. This view corresponds to the republican notion – that law is not an external infringement, but constitutive for freedom – even more than the account in *Not Just Deserts*. In our view, however, the problem is that the consequentialist component "gets lost", respectively moves to a "level below". If legally guaranteed options cannot be weighed up any more, what is the subject of promotion then? If legitimate law does not infringe liberty and one follows a consequentialist approach to determine what the legitimate law should be, one has to weigh up which *factual* options *have to be legally guaranteed*, which means law still depends on a result of weighing up.

[6] Pettit's later theory operates with the criterion of *freedom as non-domination*. A person dominates another "to the extent that (1) [she has] the capacity to interfere, (2) on an arbitrary basis, (3) in certain choices that the other is in a position to make". (R, 52) The interference is arbitrary, if it is not "forced to track the interests and ideas of the person suffering the interference […] at least the relevant ones" (R, 55). Relevant interests and ideas are those the person has in common with others, i. e. that are not "sectional or factional" (R, 56). Freedom as non-domination exists, if at least one premise is not fulfilled. It is, like dominion, the "supreme, perhaps even the unique political value" (NJD, 67).

could realize in solitude, but only of those ones you can realize in a society with other people around (NJD, 57, 63). On the one hand, this reduces the options available; on the other hand, freedom consists against this background not merely of factual, but of legally protected options (NJD, 57).

In the following, we will analyze whether the target of a promotion of 'dominion' fits the three mentioned desiderata. We will concentrate on two issues, namely if the promotion of dominion can guarantee rights and if it is a target clear and distinct enough to evaluate measures in the criminal justice system.

7.6 Promotion of 'Dominion' and Rights

As we have seen, Braithwaite and Pettit refute the preventionist and utilitarian criteria for not respecting rights. Does, however, the target of 'dominion'-promotion satisfy this demand itself? Does it provide a stable allocation of rights, i. e. will it not motivate the actors of the criminal justice system to breach those rights on occasion? The authors give an affirmative answer to this question. If 'dominion' is to be legally promoted, "then certain negative liberties must certainly be legally protected" (NJD, 71). In every case it "seems certain" (NJD, 72) that the uncontroversial rights associated with criminal justice will be included and that this allocation will be stable (NJD, 72–73). Although one could assume that target to be "liable to motivate the occasional breach of such legal rights – say, the occasional imposition of penalties on the innocent" (NJD, 72), the authors think that the "concern to promote dominion" requires actors "to internalize a commitment to promote the rights of the innocent and many other uncontroversial rights" (NJD, 72).

The point is based on the assumption that actors who take the supreme goal of promoting dominion seriously, will have to make a commitment "to tie their hands in regard to how individuals should be treated" (NJD, 73). They will "not *always* be determined by direct consideration of the target."[7] It will be "self-defeating" to pursue the promotion of dominion "in an exclusively direct fashion" (NJD, 74).

Indeed, we find this rather puzzling. As mentioned above, Braithwaite and Pettit have followed, among others, a consequentialist approach, because a purely constraint-based theory cannot provide a criterion for a comprehensive theory. Here, however, the authors introduce constraints to their theory. Which status do these constraints have? One option would be that they are values *in their own right*. For example, the stabilizing-desideratum could be applied as an independent criterion besides the target of dominion-promoting and thus guaranteeing uncontroversial rights. However, this possibility is excluded by the fact that the authors follow a monist value theory, namely a consequentialist one. The remaining option, therefore, is that constraints are *derivative* values, derived from the main consequentialist target. This could be a case of the above mentioned requirement for consequentialist theories to make room for constraints "at some less fundamental

[7] NJD, 74 (our accentuation).

level" (NJD, 32). Unfortunately, we have some doubts that a combination of a supreme monistic target and less fundamental constrains will work properly, i. e. that those constraints can be taken seriously.

Saying that actors should not "always" pursue the promotion of 'dominion' directly implies (not logically, but pragmatically) that they have to pursue it *sometimes*. As mentioned above, the promotion of 'dominion' is not only a criterion for evaluating systems, but also for evaluating particular measures. The obvious question, however, is, *when* do the actors have to apply the criterion directly. As there is one supreme value – the promotion of 'dominion' – and the constraints are derived from this value, the only answer can be that one has to analyze the supreme value to know when to apply it directly and when not. This leads to a paradox, as analyzing what the consequentialist target requires and acting accordingly *means* promoting the value. The realized correct decision of whether to directly promote a value or not, would be itself an act of value promotion. Therefore the value could *only* be promoted directly.

As the promotion of 'dominion' requires maximizing the overall dominion, and 'dominion' consists of legally granted options (rights), one would have to weigh up those options in such a way that the overall amount of those options is maximized. Due to a comprehensive comparability of all the options, no option can be guaranteed, as it might be outweighed by (a sum of) other options. An option that can be outweighed, however, does not fully deserve the title "right" (it is not a constraint, as it does not obtain independently of other things).

7.7 The Promotion of 'Dominion': Too Vague and Too Demanding?

Our second remark is in many ways interrelated with the first one, and it probably even follows from it. As we have shown, the aim of promoting 'dominion' can hardly tolerate constraints that could tell actors what to do in particular cases definitely and independently from other circumstances. This means that actors of the criminal justice system have to interpret the supreme goal directly in every case in order to know what they have to do. Does, however, the goal of 'dominion'-promotion provide them with a clear account on this question? Unfortunately, we are doubtful that it will do much better in this regard than utilitarianism, which Braithwaite and Pettit refuted, among other things, for its extreme vagueness (NJD, 52).

'Dominion' consists of different options – basically of all the options a person can choose to realize (within society, according to the republican modification). There are options differing in type and importance, falling under such classical liberties as those of "expression, movement, [...] association and [...] ownership" (NJD, 62). Moreover, these options are not only postulated as all being valuable and requiring to be weighted up *somehow*. Their congregation in one single value implies that there must be a *reasonable scale* to weigh them up.[8] Most of the named options, however, are incommensurable.[9] There is no rational account of which one is more worth than the other. There is no certain rule do decide between different options containing different "packages" of values. So how could, say, a policeman or judge realize what the goal of 'dominion'-promotion expects him to do in actual cases?[10]

7.7.1 The Application of the Goal in Practice: Determining Sentences as an Example

In order to facilitate these kinds of judgments, Braithwaite and Pettit formulate four "presumptions the republican stance supports" which should serve as "middle-range principles for interpreting the abstract goal endorsed by republicans: the promotion of dominion" (NJD, 87). We will now look at what these principles are and (as an exemplary case) how they help the authorities in determining criminal sentences.[11]

The first principle is that of "parsimony". It suggests "less rather than more criminal activity" (NJD, 87). In particular, criminal authorities have to be reserved with punishment, until they can prove that a harder sanction will increase 'dominion' overall.[12] The second principle is defined as "checking of power". It stands for the necessity to control the criminal justice authorities in such a way that they do not abuse their jurisdictional power to infringe individual rights of citizens in an arbitrary way.[13] Furthermore, there is a third presumption consisting in the "reprobation" of crimes. It expresses the goal to instill a moral feeling of shamefulness and disapproval towards crimes. The reason for such an education is to reduce the

[8] Although we know that the authors have abstained from commenting on the question of comparability of different values (NJD, 68), we nevertheless point at this aspect as we regard it important for the persuasiveness of the theory, especially in regard of the "uncontroversial" desideratum that the authors themselves mention when discussing utilitarianism (NJD, 52).

[9] We refer to the definition of incommensurability in Raz (1986, 322): "A and B are incommensurate if it is neither true that one is better than the other nor true that they are of equal value."

[10] We assume that this problem led the authors to reintroduce constraints to their theory.

[11] NJD, 101–106. Pettit formulates similar principles in *Republican Theory and Criminal Punishment*. Those principles are that of *rectification* (negative effects of crime should be undone as far as possible), that of *recognition* (offender has to understand he has done wrong), that of *recompense* (restitution of the losses caused by the crime) and that of *reassurance* (community must regain the assurance that their non-domination is safe), cf. Pettit (1997, 72–77).

[12] The reason for this is that any "act of criminalization, surveillance, investigation, or arrest, any prosecution or punishment does immediate and unquestionable damage" to the 'dominion' of the person to punish, whereas "the benefits promised by the initiative are almost always of a distant and probabilistic character" (NJD, 87).

[13] This principle is derived from the subjective component of dominion (NJD, 64–65). According to the second premise of dominion, citizens must know that the state does not want to infringe their rights but to act in a justified way (NJD, 88).

crime-rate and to make people understand the way the criminal justice system works (NJD, 89). A fourth principle is that of "reintegration", which serves to undo the 'dominion' losses of the offender (whose 'dominion' is declined by the punishment), but especially of the victim (whose 'dominion' is even more declined by the act of crime. NJD, 91).

As far as punishment is concerned, Braithwaite and Pettit distinguish three different categories: First, punishment directed against the offender's "property", second, punishment that restricts the offender's "province" (roughly the liberty of movement and action) and lastly punishment that invades the offender's "person". The first group includes fines, restitution and seizure of assets, the second imprisonment and community service and the third capital punishment, corporal punishment, mutilation und torture (NJD, 102). Now we look at how the four principles are applied.

According to 'parsimony' now, the third category should be completely prohibited because physical penalties like execution and torture invade 'dominion' in a more extensive way than interferences concerning 'property' and 'province' (NJD, 102). Furthermore, it can be derived from the first principle that there is a preference of the first category towards the second one: Restitutions do not infringe the offender's 'dominion' as much as an imprisonment (NJD, 103). The principles of 'reintegration' and 'reprobation' lead to the same preferences, because restitution brings the offender into a relation to the victim mediating him the effect of moral education needed for his resocialization and his understanding of the reprobation values within the society (NJD, 103). The second category of punishment is only applicable when the "community has a justifiable concern to be protected from future acts of violence by the offender", or if "the offender is unable or unwilling to pay" (NJD, 104). In these cases, community service is preferable towards imprisonment: The carrying out of social work in favor of society is less invasive ('parsimony'), facilitates the offender's reintegration in the society (in contrast to his isolation in a prison) and also the victim's reintegration: The damages of the victim's 'dominion' can be compensated by the offender's community services (NJD, 104).

Another way to reduce punishments against province in consonance with parsimony to its minimum, is to grant the judges a leeway in decision-making so that they can contrive penalties which are more adjusted to the committed crime. For instance, a "violent football hooligan could report to the local police station every Saturday afternoon during the next three football seasons to wash out the cells or the police cars" (NJD, 130).

7.7.2 Do the Principles Help?

The application of these principles seems to lead to sound results in the cases discussed. The problem, however, is that these principles are not logically deduced from the target of promotion of 'dominion' but are formulated as mere "presumptions". They do not claim to provide an analytical truth about what the promotion of 'dominion' itself requires in the particular situation. If that is not the case, the question is, as in the case of constraints above, what the status of those principles is. While not being deduced from the supreme target, they cannot be principles in their own right, besides the target, too (there are no competing values). Therefore, the application of those principles depends, again, on what the main goal of dominion-promotion itself requires.

This can be shown using the example of determining sentences. The principles are vague guidelines and provide the judge with no definite information in which case he has to impose a penalty of restitution, of community service, or of imprisonment. The authors, however, could argue that this vagueness – in this case the indeterminacy of which punishment is to impose – is necessary to increase 'dominion'. It allows the judge to show mercy to an offender who can be partially exonerated due to certain circumstances. If, for example, the offender has honestly realized he had done wrong and has paid restitution, then there is no reason to infringe his liberties. Strict principles of punishment would therefore violate 'dominion'.

On the other hand, some far more determined rules of punishment would be required in other cases. If, for example, a serial killer threatens many citizens, it is obvious that the promotion of 'dominion' would clearly require him to be arrested and imprisoned. In that case, vague punishing-guidelines jeopardize the promotion of 'dominion' overall.

Notice that in both cases we refer to our more or less intuitive understanding of 'dominion' to decide whether a strict or a vague rule is necessary. The paradox result is this: The four principles are formulated to help actors "interpreting the abstract goal endorsed by republicans" (NJD, 87). But in the end, one has to interpret the abstract goal itself to determine if and how the principles are to be applied. The principles, therefore, cannot eliminate the actors' problems to evaluate which options the promotion of 'dominion' regards as preferable in different cases.

To sum up, we formulate two points of critique: First, the supreme target – the promotion of dominion – does not seem to take rights seriously and is therefore unstable. Furthermore, the target of a promotion of dominion does not give a clear account on how to weigh up different options. Particularly, the four republican presumptions cannot serve that purpose, as they neither are logically deduced from the target, nor can they be applied in their own right.

References

Pettit, P. 1997. Republican theory of criminal punishment. *Utilitas* 9: 59–79.

Raz, J. 1986. *The morality of freedom*. Oxford: Oxford University Press.

Chapter 8
Indirect Consequentialism and Moral Psychology

Anna M. Blundell, Simon Derpmann, Konstantin Schnieder, and Ricarda Geese

8.1 Introduction

If one takes a look at recent debates in philosophical ethics, one soon encounters a widespread tendency to organize complex debates into mutually exclusive factions attached to big labels such as “consequentialism”, “deontology” or “perfectionism”. Since consequentialism is one of the major traditions of ethical reasoning, it does not surprise that it is exposed to a certain amount of criticism. Consequentialism is accused of ignoring individual rights, of legitimizing the tyranny of the majority, or of ignoring moral intuitions and overcharging ordinary moral agents.

In what follows, we want to revisit one of these objections and hope to cast some doubt on whether Pettit’s account of consequentialism successfully answers it. Our argument will proceed in three steps. First, we will reconstruct what Pettit takes to be the core idea of consequentialism, defining moral rightness as the promotion of neutral value. This constitutive feature gives rise to the objection that consequentialism cannot be true to our common moral psychology, an objection that Pettit responds to in his account of consequentialism. Pettit perceives these moral-psychological facts as a potential threat to consequentialism, yet he wishes to embrace them almost without reserve. Having introduced Pettit’s indirect consequentialism, we will then ask whether indirect consequentialism is a coherent answer to this challenge and whether it is able to reflect our moral psychology adequately. We will eventually object that under indirect consequentialism moral agents have to abstract or alienate themselves from their personal projects and relations no less than under direct consequentialism. If our arguments are sound, then indirect consequentialism comes under fire from two different directions: Either it is threatened to turn into a self-effacing theory that only contains a justification of our moral practices but

A.M. Blundell • S. Derpmann (✉) • K. Schnieder • R. Geese
Philosophisches Seminar, Westfälische Wilhelms-Universität, Münster, Germany
e-mail: annamblundell@gmail.com; simon.derpmann@wwu.de;
konstantin.schnieder@uni-muenster.de; ricarda.geese@gmx.de

S. Derpmann, D.P. Schweikard (eds.), *Philip Pettit: Five Themes from his Work*,
Münster Lectures in Philosophy 1, DOI 10.1007/978-3-319-26103-4_8

is unable to guide moral agents; or it effectively guides moral agents in a revisionary way that neglects the idea of assigning special normative significance to one's own personal commitments. Our argument does not amount to a rejection of consequentialism, but rather elucidates the implications of consequentialism for common moral deliberation.[1]

8.2 Pettit's Definition of Consequentialism

What makes an option right within a particular choice, so consequentialism teaches, is that it advances the agent-neutral good more than any other possible option would. According to Pettit, consequentialism can thus be distinguished from competing theories by two defining components, namely the *promotion* of *neutral* values.

> The consequentialist says, first, that values determine rightness in the promotional way, not the honoring way. And the consequentialist says, second, that the values which determine rightness are all neutral values, not values that have distinctively relativized reference. (CP, 129)

Consequentialism requires moral agents not to "honor" certain values by not infringing them – say, by not breaking promises or not hurting anybody – but rather asks moral subjects to act so as to best promote the overall amount of values such as honesty or peace. Furthermore, according to the consequentialist picture, moral values are agent-neutral. Pettit explains that a

> value will be a neutral value, we can say, if and only if we can know what it is that is valued without knowing who the valuer is. It will be a relativized value, on the other hand […] if and only if we cannot know what it is that is valued without knowing the valuer's identity. (CP, 125)

Thus, the consequentialist supposes that what is valuable and morally required can be recognized without referring to specific positions, relations or histories of the person who passes a moral judgment. That the welfare of a friend of mine has relative value for me can be seen from the fact that I cannot grasp the specific importance of his flourishing without thinking of my relation to him. I care for him as *my* friend, and in order to do so, I must reflect my own position in the world as being *his* friend.[2] That, on the other hand, education or knowledge possess neutral value

[1] Admittedly, Pettit does not claim to incontrovertibly prove the "inescapability of consequentialism". Rather, he ultimately suggests that the political need for consequentialist deliberation and the desirability of theoretical unity in practical philosophy give us good reason to accept consequentialism in the domain of private morality, too. Since we do not want to engage in a metaphilosophical debate on different theory ideals in philosophy here, our main aspiration in this paper is to indicate that the price for theoretical unity is higher than Pettit suggests.

[2] Assigning relative value to friendship is thus different from recognizing friendship as a reflexive value. The former gives me a reason to promote the well-being of a friend of mine, the latter gives me a reason to promote instances of friendship. See Pettit (2012, 48).

can be seen from the fact that I do not have to position myself in order to approve of their promotion. It should not matter to me whose education or knowledge it is that is furthered.

A similar observation holds for the promotion of values. Should a clever lie, for example, effectively promote the overall amount of honesty in the world, it should not matter to the agent that acting consequentialistically requires him to tell a lie. Since he is committed to promoting the neutral good and lying is in that case the best way to realize that goal, he must rather approve that action to be morally right. Now, that picture of moral agency in turn provokes a severe objection against consequentialism, which we will consider in the following sections.

8.3 Moral Psychology and Indirect Consequentialism

Having stated briefly what Pettit holds to be the defining theses of consequentialism, we will now turn our attention to an objection against consequentialism that Pettit hopes to answer with his theory of indirect consequentialism. This objection criticizes that consequentialism is revisionary to our common moral psychology. As Pettit concedes, "opponents see consequentialism as culpably and fundamentally misrepresenting the moral psychology of agents." (CP, 93).

8.3.1 The Moral-Psychological Objection

More precisely, consequentialism is said by its critics to ignore that the motivations of individuals are often non-atomistic and non-moralistic and that moral agents do not usually make their decisions in a calculative way (CP, 94ff.). When human beings act, so the non-consequentialist objects, they are motivated by considerations that involve other people essentially (non-atomism), that encompass personal commitments and projects apart from the abstract demands of morality (non-moralism), and they do not constantly calculate what they should do (non-actuarialism). Since consequentialism seems to require of moral agents that they should consciously subject personal commitments and intersubjective allegiances to the demands of promoting neutral goods, those moral-psychological facts seem to give the lie to consequentialism.

Now, a consequentialist could respond to this challenge in three different ways. (i) He could simply accept that consequentialism is revisionary of our commonplace psychology, but insist that it is nonetheless the right theory. If moral philosophy tells us that some aspects of our moral practice are ill-guided, he could argue, then we need to revise our practice, not our moral theory.[3] If the consequentialist

[3] Naturally, a moral philosopher going in this direction faces a lot of different challenges, but that is beyond our main focus in this article.

wants to safeguard at least some of our moral psychology against such demands, however, he could choose between two further responses. (ii) He could enrich his theory with additional moral principles or morally relevant aspects besides the promotion of agent-neutral goods and, thus, turn his theory into some sort of hybrid or pluralistic moral theory drawing upon consequentialist as well as deontological or virtue-theoretical resources. Should he not want to transform his theory into a hybrid conception of morals, but rather rely on a single consequentialist principle, then he could (iii) try to show how the moral-psychological facts can be allowed in a purely consequentialist morality.[4] Rejecting both the first and the second strategy, Pettit argues that, on the one hand, the way ordinary human beings are in their psychological make-up "is how they should be allowed to remain under any plausible moral theory" (CP, 102), but that on the other hand moral rightness consists in the promotion of neutral values.[5] By trying to show that agents have consequentialist reasons to follow non-consequentialist motives, Pettit is, therefore, taking the third possible response strategy to reconcile consequentialism with our moral psychology through his specific brand of "virtual" or "indirect consequentialism" (CP, 93 ff.). Understanding what 'going indirect' means on Pettit's account and questioning whether that strategy is able to successfully evade moral revisionism without running into new problems is the subject of the following sections.

8.3.2 The Indirect Consequentialist Replies

According to Pettit, any strategy of indirection of moral deliberation and decision-making is characterized by three aspects: "First, it offloads active control to a modular pilot that operates more or less autonomously, like the modularized skill revealed in typing and in tying your shoelaces; second, it preserves virtual, standby control by keeping the agent ready to deliberate on a need-for-deliberation basis; and third, it outsources the trigger that prompts deliberation, letting external red lights dictate whether to reclaim active control or not." (Pettit 2012, 47) With specific regard to consequentialism, he says that

> [u]nder an indirect strategy, deliberation over the enactable options will be restricted so as to make room for the operation of suitable predispositions: suitable plans or policies, motives or habits or traits, or commitments to social practices. The right is always a

[4] More precisely, one could, again, distinguish between three different ways of making pure consequentialism compatible with the moral-psychological facts (Scheffler 1992, 17 ff.). A pure consequentialist could, firstly, narrow the *scope* of his moral theory and, thus, exempt certain areas of human action from the demands of his theory. Secondly, he could challenge the *authority* of morality and so concede that it is sometimes rational for people to ignore the demands of morality. Or he could, thirdly, qualify the way moral considerations must enter into the *deliberation* of moral agents. In this paper, we will not discuss all three possible options and instead concentrate our discussion on Pettit's answer to the moral-psychological objection.

[5] See also Pettit (CP, 161): "Any decent moral theory must enable defenders to sustain the ordinary sort of moral psychology described […]."

> promotional function of the good but in the case of action that function may have an indirect character. (Pettit 2012, 47)[6]

Here we can see what is indirect about Pettit's consequentialism and how this may help him overcome the challenge from common moral psychology. He specifies our mode of moral deliberation so as to allow us to be guided by non-consequentialist motives. Since the right is nonetheless determined by the good, Pettit's moral philosophy can still be understood as consequentialist. Pettit illustrates this in the example of someone's reasons to help a friend move his apartment. From a perspective of the promotion of neutral values, one should not help others because they are one's friends, but because helping them leads to the most valuable results. Nonetheless, indirect consequentialism may allow a person to give special consideration to her friends. On that picture, she may do so not because she believes to be specially obliged towards them, but because she holds the belief that allowing herself to be moved by partial considerations of friendship promotes the greatest value. However, she will not reflect this consideration of moral neutral value while she helps her friend, because this would impair some attitudes that are constitutive of her relation towards her friend. Rather, indirect consequentialism calls for such critical reflection of friendship (and other relativized motives) only when they lead to nonoptimific results, as in the case of helping her friend move a body. Reconstructed in that way, the structure of Pettit's indirect consequentialism bears resemblance to what some pragmatists have called the default-and-challenge model of justification. Under normal circumstances, so Pettit tells us, an agent may offload active control over his actions and just let himself be carried by "natural affection" (Pettit 2012, 47) without going over any explicit consequentialist deliberation. Only when the "red lights", that "will go on, ideally, just when the cause of the neutral good is likely to be jeopardized by letting sensibility rule on its own" (Pettit 2012, 46), indicate the need for deliberation, must he discharge those predispositions and reclaim active control over his actions. Even though he does not constantly exercise active control, then, such an agent never loses a form of "virtual or standby control" over his action, "for he is ready to intervene on a need-for-action basis" (Pettit 2012, 46). In order to make room for the operation of "suitable predispositions" (Pettit 2012, 47), such as friendship or love, indirect consequentialism restricts the need for moral deliberation over the right option to cases where unreflected motivation is unlikely to promote agent-neutral goods.[7] That way, a

[6] As indicated before, Pettit attempts to counter the moral-psychological objection by specifying the mode of *deliberation* his theory calls for. With this move, he especially wards off an objection that has, since Bernard Williams, become popular under the catchphrase "one thought too many" (Williams 1981, 18). While Williams's famous critique seems to rest on the presupposition that a consequentialist must hold the moral norm to be the "motivating thought" (ibid.) of an agent, Pettit's theory explicitly curbs the deliberative role of moral considerations by limiting direct consequentialist deliberation to morally alarming situations.

[7] Of course, the causal origin of an action is not morally relevant in itself. It is only morally relevant insofar as the causal origin reflects certain behavioral patterns that are relevant to the promotion of the neutral good.

consequentialist can evaluate and justify certain predispositions with regard to the expected good they produce when they make it more likely that the good will be promoted, e.g. because they encourage the best actions or because they guarantee interpersonal reliability. So if a consequentialist agent chooses to help a friend, what makes his action right is not determined by the content of his choice alone, but also by the instrumentally valuable predisposition manifested in his action. As long as such predispositions do not result in sub-optimific consequences, indirect consequentialism allows people to let themselves be guided by their ordinary moral motivations.

8.4 Recurring Objections

Because it acknowledges the beneficial character of our non-consequentialist moral psychology, indirect consequentialism can supposedly make room for non-consequentialist deliberation. Still, this sort of indirect consequentialism provokes a different objection, which we discuss in this section. By going indirect Pettit might be able to evade the moral-psychological objection, but he seemingly exposes his theory to another problem. For, an indirect moral theory such as Pettit's appears to invite the objection of turning consequentialism into what Parfit calls a 'self-defeating theory'. One may suspect that the indirect consequentialist believes that the world will be best in terms of achieving the promotion of agent-neutral goods, when moral agents are genuinely motivated by agent-relative (i.e. non-consequentialist) considerations, such as friendship or fidelity, and do not make the promotion of neutral goods their conscious objective. Since indirect consequentialism would apparently give moral philosophers reasons to ban or hide consequentialism from the moral deliberation of agents, it would ultimately render consequentialism self-effacing.

8.4.1 Is Indirect Consequentialism Self-effacing?

The essential step towards answering to this objection is Pettit's differentiation between two modes of control, active and virtual. According to Pettit, an agent does renounce active control over his behavior when he lets himself be guided by mere habits etc., but still retains virtual control as long as he could intervene into his actions. This picture supposedly enables Pettit to prevent his theory from becoming self-effacing, since under indirect consequentialism agents remain true consequentialists, who sometimes explicitly evaluate their actions by reference to the consequentialist criterion of rightness.[8] While they retain virtual control and stay "ready

[8] Pettit explicitly builds indirection into his account of consequentialism to answer this challenge. See Pettit (2012, 45): "Does it amount to the sort of self-effacing consequentialism in which the

to deliberate", they cannot simply ignore consequentialist reasons when being guided by habits, intuitions etc., but rather make reference to consequentialism once the external warning signs hint at a need for deliberation. Equipped with a suitable pattern of predispositions that help to promote neutral goods, then, they need not constantly evaluate e.g. their non-atomistic or non-moralistic motives, but rather go over such explicitly consequentialist deliberation only if their habitual motivations might not be optimific.

Yet, it is hard to imagine how the 'red-lights' or the 'auto-pilot' work and who is in charge of keeping an eye on the red light while an agent has suspended active control and comprehensive deliberation. There are two options, each confronted with a specific problem. On the one hand, every agent may supervise his acting from mere predispositions by himself and endorse this mode of action for consequentialist reasons until he might hit a point at which morality demands a change of behavior. In that case, however, this type of consequentialism is only semi-virtual, since agents would constantly think of ways of justifying their actions and could, therefore, leave guidance to their predispositions within narrow limits only. Such agents could, thus, only help their friends with due reservation, as they would constantly ask how far they could go down that path before no longer having any consequentialist justification for their actions. Under this image, then, agents are forced to split themselves into (seemingly) spontaneous, amicable, or loving persons on the one hand and vigilant backdoor consequentialists on the other hand.[9] The second option, the one closer to a social practice of default-and-challenge, poses a quite similar problem for Pettit's theory. While in the default-and-challenge model, the moral agent is not divided into a non-consequentialist actor and a consequentialist deliberator with regard to his own actions, he now faces the problem that he may act and deliberate as a non-consequentialist in his own daily life, but that he has to judge the actions of others by a consequentialist measure. Splitting his personality would, therefore, reoccur even under the second image.

Both these observations show that Pettit's indirect consequentialism faces a serious problem in attempting to both maintain the inclusion of nonconsequentialist deliberation in order not to become revisionary with respect to our moral psychology, as well as the ultimate reliance on consequentialist moral reasons in order not to become self-effacing. This problem will be the main topic of the remainder.

agent eliminates the possibility of deliberating over consequences (Parfit 1984)? No, because I offload only active control, not control period."

[9] As part of her critique of "moral sainthood" Susan Wolf (1982, 429) raises a related objection against consequentialist theories, claiming that a "limited and carefully monitored allotment of time to be devoted to the pursuit of nonmoral interests […] would make a person a better contributor to the general welfare." Although her critique deviates from our argument in some respects, Wolf also emphasizes that "the need to monitor will restrict not only the extent but also the quality of one's attachment to these interests and traits" (ibid.).

8.4.2 Psychology Revisited: Reasons and Motives

If Pettit does not want to turn his consequentialism into a self-defeating theory, then consequentialist reasons cannot be a mere philosopher's description of how moral agents decide even if *they themselves* deliberate on wholly other terms. What is needed for indirect consequentialism to be an action-guiding normative theory is an agent's higher-level commitment to consequentialism. If he remains in virtual control over his behavior and actualizes it once he judges his undeliberated practice to be sub-optimific, then he obviously shares a basic commitment to optimize his action in a consequentialist sense. This is "the consequentialist currency to which they pay allegiance." (CP, 158). Without questioning the plausible assumption that moral agents sometimes suspend comprehensive deliberation and active control over their behavior, we want to argue that even such a restrictive consequentialism does not pay adequate regard to the way ordinary moral agents actually do think about moral questions. In particular, we want to cast doubt on the claim that what Pettit calls "predispositions" such as friendship or love are really some sort of modular pilots that enjoy only derived justification. Against that position, we will, firstly, ask whether such a predisposition is a "modular pilot that operates more or less autonomously" (Pettit 2012, 47) and, secondly, revive the objection that our common-place morality is not *only* concerned with the promotion of agent-neutral goods, but also permits agents to understand their own personal projects and relations *as reasons*. While non-consequentialist predispositions can be allowed within indirect consequentialism, they only enjoy derived or instrumental justification and do not provide reasons for actions by themselves. After all, they are only *motivations* that operate "more or less autonomously".

However, that picture misrepresents the way such non-consequentialist aspects enter into our moral self-guidance in two ways. First of all, it probably strikes ordinary people as absurd if one compares their interaction with a friend or a loved one to the "modularized skill revealed in typing and in tying your shoelaces" (Pettit 2012, 47). When a person helps a friend, for example, he often does not simply let an autonomous cognitive process guide his behavior, but rather consciously decides how to act. And when he decides, his friendship, love or promises do not guide his action as instrumentally justified motivations only, but as self-standing normative reasons for or against certain actions. Such a person would, then, act as a friend not because he knew that his behavior was consequentialistically justified or that he could give such a justification if necessary, but because to him his friendship provided a morally relevant reason in itself. Of course, that reason might sometimes be overridden or trumped by neutral or other moral demands. However, that sort of moral conflict is possible only if he recognizes that friendship enters into moral deliberation as a genuine reason to be weighed against others.

Pettit's inclusion of non-consequentialist motivation seems adequate for those cases, in which the right action demands a suspension of consequentialist reasoning, because we are more likely to do the right thing intuitively or habitually. Yet, there is a difference between the claim that sometimes we should act *not following* *consequentialist deliberation* and the claim that sometimes we should act *following non-consequentialist deliberation*. Instances of the first kind seem to be covered by Pettit's model of indirect consequentialism, while instances of the second kind do not. Our moral psychology, however, demands motivations of the second kind which are hard to conceive under indirect consequentialism.

Pettit would, of course, allow agents to care for such agent-relative goods; yet, he would only do so since on his account that care could be neutrally justified. In the end, that amounts to erasing agent-relative reasons from the range of normatively relevant considerations while reintroducing them for consequentialist reasons on a secondary level of normatively irrelevant motivations. Since Pettit is, as we have seen, not willing to accept the revisionary implication of consequentialism, or enrich his theory with non-consequentialist principles, this dissent with Pettit ultimately turns on the truth of consequentialism, be it indirect or direct. Respecting our moral psychology and recognizing the complexities of moral agency, so we have argued, forces moral philosophy to acknowledge that we do not solely pay allegiance to the consequentialist currency, but rather worship competing moral demands.[10]

8.5 Conclusion

In this article, we have portrayed Pettit's version of 'indirect consequentialism' and discussed various objections to it. In the end our own criticism looks like a revenant of those objections that Pettit claims to have laid to rest by 'going indirect'. Either a moral agent must divide himself into a consequentialist supervisor of his own action and his spontaneous personality, which seems psychologically as implausible as full-blown actuarialism; or he is allowed to simply ban consequentialist supervision from his moral deliberation, and consequentialism, thus, becomes a self-defeating theory.[11] The fundamental challenge to consequentialism builds on the demand of taking our moral psychology seriously. If our argument is convincing, Pettit should account for the fact that people are not only *motivated* non-moralistically or non-atomistically but that they think that considerations of this kind give them *genuine reasons for action*.

To stick with Pettit's metaphor, then, we would suggest that moral agents, who are in control of their behavior, not only pay attention to consequentialist "red lights" but rather stand at the crossroads of multiple and potentially conflicting

[10] To be sure, this thesis does not commit us to the *normative* claim that non-consequentialist reasons should be privileged or cannot be outweighed by consequentialist reasons.

[11] Jonathan Dancy (1993, 236) poses a dilemma of this sort to consequentialism: "There is an awkward dilemma here. The first horn is one in which the admission of the agent-relative is a mere sham or pretence, because the way in which we are allowed to find value in love and friendship is not one which fits any value that those states actually have […]. On the second horn we allow that there really is the sort of value that we find in those states of partiality, at the cost of preventing ourselves from reasserting the dependence of such value on neutral value."

moral demands. Due to these considerations, it seems that the type of consequentialism discussed here is not particularly suited as a means of guidance in questions of morality, but rather (if that) a description of the way moral decisions are made on an everyday basis with an attempt to integrate consequentialism. We agree with Pettit that moral theory should provide guidance for ordinary human agents, while allowing them to preserve their fundamental psychological make-up. However, if our arguments are convincing, indirect consequentialism is unable to meet both these requirements.

References

Dancy, J. 1993. *Moral reasons*. Oxford: Wiley-Blackwell.

Pettit, P. 2012. The inescapability of consequentialism. In *Luck, value and commitment: Themes from the ethics of Bernard Williams*, ed. U. Heuer and G. Lang, 41–70. Oxford: Oxford University Press.

Scheffler, S. 1992. *Human morality*. Oxford: Oxford University Press.

Williams, B. 1981. Persons, character and morality. In *Moral luck*, ed. B. Williams, 1–19. Cambridge: Cambridge University Press.

Wolf, S. 1982. Moral saints. *The Journal of Philosophy* 79: 419–439.

Chapter 9
What Is the Foundation of Pettit's Non-redundant Realism About Group Agents?

Dominik Düber, Nadine Mooren, and Tim Rojek

9.1 Introduction

Since the publication of his book *The Common Mind* in 1993, Philip Pettit is dealing with problems of social philosophy in general and with philosophical problems of an adequate social ontology in particular. In *The Common Mind*, Pettit divides the field of social ontology into two areas of debate. With the beginning of the new century, Pettit added a third field of debate to his map of social ontology (Pettit and Schweikard (2006), 35–37). He maintains that answers in one field are compatible with different answers on another field, so the areas of debate are at least partially independent. Roughly since 2001,[1] Pettit is working, partly in cooperation with Christian List, on the third field of social ontology which is called the *singularism* debate.[2] In this field of debate "[t]he issue is how far people can unite to form intentionally minded agents of a group kind: agents that constitute institutional persons with minds of their own" (JD, 291).

In this paper, main interest is in the development and theoretical moves Pettit makes in this debate. So our paper is more or less the attempt to reconstruct his position in a clear way and to make sure that we got it right. We will refer to the positions developed in the last 10 years as Pettit's positions. We do this as a matter of simplicity and not to deemphasize the contribution of others, e.g. Christian List.

[1] We take the following papers as starting point for Pettit dealing with the third issue: Pettit (2001a, b).

[2] The term *singularism* was introduced in the debates of social ontology by Margret Gilbert (1989).

D. Düber
Kolleg-Forschergruppe Normenbegründung in Medizinethik und Biopolitik,
Westfälische Wilhelms-Universität, Münster, Germany
e-mail: dominik.dueber@uni-muenster.de

N. Mooren (✉) • T. Rojek
Philosophisches Seminar, Westfälische Wilhelms-Universität, Münster, Germany
e-mail: nadinemooren@gmx.de; t.rojek@web.de

S. Derpmann, D.P. Schweikard (eds.), *Philip Pettit: Five Themes from his Work*,
Münster Lectures in Philosophy 1, DOI 10.1007/978-3-319-26103-4_9

9.2 Mapping the Debate

In the book *Group Agency*, which was published in 2011, Pettit unifies the work of the past 10 years, combining the different fields of work belonging to the third debate. The book is divided into three parts, but we will mainly deal with the introduction and the first part. In this part of the book Pettit aims at defending "the logical possibility of group agents" (GA, vii) or "the possibility of group agents" (GA, 2). Our main intention is to reconstruct the argument and aim of the first part and to understand who the addressee of this account is.

In the introduction Pettit differentiates between different accounts concerning the question whether our every-day talk about groups intending something or doing something as groups, should be understood as an adequate language-game or if we – as philosophers – have to avoid it because it has no socio-ontological justification. Pettit wants to defend the position that at least parts of our every-day talk about group agents can be justified, even if it is misconceived in other cases.

Since such an account deviates from everyday talk on some occasions, whereas it reconciles it on other occasions, it follows that *neither* the defender of group agency, *nor* the defender of a singularist position is fully in accordance with our common-sense view and talk about groups (Schweikard 2011, 317–318). So the burden of proof seems to be distributed equally between the two parties.

Pettit himself defends a position that he calls "non-redundant realism". He differentiates between three types of enemies of a non-redundant realist position: two versions of *eliminativism*[3] and a *thin, redundant realism* (GA, 7). The first kind of eliminativism takes *every* utterance involving groups as intending or acting subjects as merely metaphorical. We call this position *metaphorical eliminativism* from now on. The second version of eliminativism is an error theoretic approach. It takes group-level talk as non-metaphorical, but says that every description of a group as acting or intending something is wrong. We call this position *error-theory eliminativism*. The third and last opponent is thin, redundant realism, which says that group-level talk is neither metaphorical nor wrong, but that every instance of a group action or intention is readily reducible to individual-level talk. We will refer to this position as *redundant realism*. In order to gain an umbrella term covering both versions of eliminativism and redundant realism, we will use the term *singularism*, a term Pettit does not make use of in *Group Agency*.[4]

On the other hand, Pettit wants to demarcate his position from another strand of thought he calls "emergentism" or "animation theory" (GA, 9). Pettit describes such a position as explaining group agency as "[…] the product of an equally mysterious, organicist force" (ibid.) as the *vis vitalis* of early biological vitalists.

[3] In a short remark Pettit says that it is possible to "re-describe" eliminativism "more positively as 'singularism' […]. Singularism asserts that there are only individual agents and that any talk of group agents is either metaphorical or wrong." (GA, 3). So Pettit's old term for the third debate has now become a re-description. In the following, we take singularism as an umbrella term covering both versions of eliminativism and redundant realism.

[4] Except in mentioning the position of Margret Gilbert (GA, 3 and 74).

Pettit calls the position he seeks to defend non-redundant realism:

> In virtue of our claim that group agents not only exist but that talk about them is not readily reducible to talk about individual agents, we may describe our account as not just a realist, but a non-redundant realist theory. (GA, 6)

Unfortunately, this labeling is slightly misleading, since emergentist accounts are also non-redundant realist. More precisely, his position would have to be called non-emergentist, non-redundant realism. For the sake of simplicity, we will stick to Pettit's labels and use non-redundant realism as an anti-singularistic *and* anti-emergentistic position.

9.3 The Relationship of Ontology and Methodology

Pettit thinks that the different forms of singularism are the majority position which "seems supported by a methodological conviction at the heart of much of economics and the social sciences" (GA, 3). This 'methodological conviction' is identified by Pettit as the "methodological individualism of the philosopher Karl Popper" (ibid.) and others. But Pettit attempts to show that non-redundant realism in social ontology is compatible with individualism in methodology, although most of the methodological individualists think that it is not. Whereas singularism is committed to methodological individualism and emergentist realism is committed to methodological anti-individualism, the historical novelty of Pettit's position (i.e. non-redundant realism) consists in developing a version of realism that is compatible with methodological individualism.

Methodological individualism is understood by Pettit as the methodological rule "that economic and social explanations should resist any appeal to psychologically mysterious social forces" (GA, 3). This definition seems to be arbitrary for at least two reasons.

(i) The first reason is that the postulation of 'psychologically mysterious social forces' seems to be an ontological and not a methodological point. *This leads us to our first question*: we do not see how Pettit distinguishes between methodology and ontology. If Pettit distinguishes between the methodological and the ontological level, it may be possible to combine different approaches on both levels. This seems to be what Pettit has in mind, since he says that he can combine a position in social ontology, i.e. non-redundant realism about group agents, with a position in methodology, i.e. methodological individualism. Whereas, according to Pettit, Popper and Hayek thought methodological individualism to be committed to eliminativism, Pettit now maintains that eliminativism implies individualism, but not the other way round, since individualism is compatible with both, singularism and non-redundant realism. This distinction seems to presuppose a clear distinction between methodology and ontology we do not find in Pettit's book. Contrary to this, the way Pettit introduces individualism, it seems to be an ontological position, since research in psychology determines which entities count as mysterious forces and which do not. This finds further support in Pettit's claim that

> [i]ndividualism is to the social sciences and economics what physicalism is to biology and psychology. Physicalism is the view that biological explanations should not appeal to any physically or chemically mysterious life force […] and that psychological explanations should not appeal to any physically or biologically mysterious source of mentality. (GA, 3)

For us, this sounds very much like an ontological statement, and Pettit's description of physicalism in *Joining the Dots* underlines this finding:

> Physicalists hold that everything in existence is constituted in some way out of physical materials and that all the laws and regularities that hold in the actual world are fixed by physical laws and regularities. 'Physical' might just be taken to mean 'microphysical' or 'subatomic'; taking it in this way would avoid some troublesome ambiguities (JD, 216)

Even if we lack a precise definition of what ontology is, those statements sound like statements about the make-up of our world, which seems to be the classical realm of ontology. But if the distinction between methodology and ontology is blurred in this way, we do not see how Pettit can make use of this distinction in order to defend the compatibility of a social ontological and a methodological position – but this seems to be the central novelty of non-redundant realism.[5]

(ii) The second reason for finding the definition of individualism arbitrary is that in this form the definition seems to be tendentious or biased *against* the methodological non-individualist, because no non-individualist would accept the ascription of postulating something 'mysterious'. Furthermore, Pettit's criterion for the division between methodological individualists and non-individualists is not very clear: "What counts as psychologically mysterious may require further explication, of course, but we abstract from that issue here, taking it to be something that *research in psychology* can adjudicate" (GA, 3; our italics).[6] If this is the criterion for dividing between methodological permissible explanations, i.e. the methodological individualistic ones and the explanations which presuppose non-individualism, there are two further problems.

The *first one* is that it is – at the moment – not possible to decide if Pettit's theory is using only explanatory resources which are compatible with contemporary

[5] In the discussion, Pettit contended making use of a misleading term in calling his position 'methodological', since it indeed is an ontological position and he wants to tie his notion of individualism to the notion introduced by Karl Popper. But this commitment is not only misleading – Pettit explicitly speaks of "a methodological conviction at the heart of much of economics and the social sciences" (GA, 3) – but causes further problems. Popper's position is not a merely ontological position but makes use of a mixture of ontological and normative statements that are intertwined and connot be easily divided. But for an adequate discussion in social philosophy, it is necessary to keep ontological and normative statements apart, cf. Taylor (1989). With commiting himself to Popper's individualism, Pettit would confound these distinctions that he himself makes use of at other occasions, cf. CM and Pettit and Schweikard (2006). Furthermore, Popper's normative position would belong to the individualism-debate (see JD, 287 ff.) and is not part of the project of *Group Agency*.

[6] Beyond this, we do not find a positive characterization of methodological non-individualism in *Group Agency*.

research in psychology. In his book, Pettit does not designate that his resources – many of them stemming from rational and social choice theory – are based on contemporary psychology.

The *other problem* is that he binds the possibility to decide between mysterious and non-mysterious entities to the investigations of psychology. We would expect a positive argument, why psychology is the discipline that decides on legitimate entities. Our initial expectation is that these questions belong to the theory or philosophy of science. Furthermore, with an eye on Pettit's own methods, the direction of fit seems to be the other way round: even if research in psychology reveals that most actors in our world do not fulfill the standards of rationality developed by Pettit and rational choice theory, it does not seem as if Pettit would be inclined to give them up (GA, 24). The reason for this might be that it is a normative task to decide what counts as rational action.[7]

9.4 From Simple Agential Systems…

In chapter one, Pettit sets out to defend the logical possibility of group agents by introducing "basic conditions of agency" (GA, 19). This theory of agency provides the conceptual resources that have to be explicated before one tries to approach "*groups* as agents" (GA, 31; our emphasis). Pettit starts by imagining a simple robotic system (GA, 19) that is meant to serve as a "paradigmatic agent" (GA, 21), realizing features of agency only in limited scope, before he then progresses to expound more complex and more resourceful agential capacities (like the capability of reasoning). According to Pettit, there are three features that present the *core idea* of agency and define a system as an agent. Such a system is described as follows:

> *First feature.* It has representational states that depict how things are in the environment.
>
> *Second feature.* It has motivational states that specify how it requires things to be in the environment.
>
> *Third feature.* It has the capacity to process its representational and motivational states, leading it to intervene suitably in the environment whenever that environment fails to match a motivating specification. (GA, 20)

Pettit calls representational and motivational states intentional states (GA, 21).

[7] In the discussion, Pettit suggested that with 'psychology' he refers to the concept of 'folk psychology'. This answer stands in tension with the position in the book, where Pettit explicitly refers to "research in psychology" and does not make us of the term 'folk psychology' (neither in this context nor anywhere else in the book). Furthermore, this interpretation faces certain follow-up difficulties, e.g. (i) it becomes less clear, which entities count as mysterious, since he would have to accept all kinds of explanations which are accepted by ordinary people, even though they vary over time and may include instances of "organicist metaphors" (GA, 9) that Pettit would want to discard – at least as long it is contended that the organicist explanations supervene on individual attitudes but are not readily reducible to them. Additionally, (ii) the concept of 'folk psychology' might get in tension with Pettit's anti-reductionist account on different fields of philosophy.

He concludes the discussion of agency by describing different dimensions of standards of rationality. These standards – that do not need to be mentioned in detail for our purposes – have to be “satisfied at some minimal level if a system is to count as an agent at all” (GA, 24).

We do not want to challenge the functionalist approach that characterizes Pettit’s concept of agency but take it for granted in what follows. This approach provides the initial conditions to get from simple agents to group agents.

9.5 …to Group Agents

The question Pettit addresses in the second chapter of *Group Agency* is whether the conditions of agency developed in chapter one can be satisfied only by individuals or whether it is possible for groups to behave in such a way that they fulfill the conditions of agency so that it makes sense to speak of them as agents in their own right. This amounts to the question whether jointly intentional actions of groups are possible. For this purpose, Pettit takes the following to be the core question: “How could a multi-member group move from the distinct and possibly conflicting intentional attitudes of its members to a single system of such attitudes endorsed by the group as a whole?” (GA, 42)

Hence, Pettit takes it to be central for group agency to get from individual intentional attitudes, i.e. representational states and motivational states (GA, 21), to such intentional attitudes which are held by a group. For this purpose, so it seems, Pettit introduces the idea of an aggregation function. Such a function serves the purpose of deriving collective output, i.e. group attitudes, from individual input, i.e. individual attitudes (GA, 48). *This leads to our next question*: Pettit defines a group agent as “a group that exhibits the *three* features of agency” (GA, 32, our italics), which includes the third feature, i.e. the capacity to process intentional attitudes. It now seems as if the task Pettit addresses in what follows only aims at reconciling the individual level and the group level regarding the first two features of agency while leaving out the third, since he takes “intentional states” to be the generic term for representational and motivational states (GA, 21).[8] This is taken up again at the beginning of the second chapter, when Pettit discusses representational and motivational attitudes only. Therefore, it seems as if the aggregational functions that are discussed in the subsequent parts only answer the question of how to get from individual intentional attitudes to intentional attitudes of groups. But this does not show what takes over the function of processing representational and motivational states on the group level, i.e. the third feature of agency.

Another reading would suggest that a group agent does not have representational and motivational states in its own right, and a suitable aggregation function takes the intentional states of individuals as input and is itself the capacity to process these

[8] An intentional attitude is always an instantiation of an intentional state (GA, 21), therefore we can treat them as similar for the purposes of this essay.

attitudes, leading to intervention in the environment. In this case, the aggregation function would be the analogon to the third feature of the individual level. But in this case, it seems as if group agents do not exhibit the first two features.

We did not find a later section in the book that addresses the questions of presenting the three features on the group level, but it seems as if this would be necessary since Part I demands to have shown the (logical) possibility of group agents (GA, vii and 2).[9]

We leave this problem aside and turn towards the role of the discursive dilemma and the majoritarian aggregation function. This is connected to the *question of the addressee of Pettit's book* we mentioned in the beginning. Looking at Pettit's work on group agency in the last 10 years, we have the impression that the function of these elements in arguing for the possibility of group agents points towards two very different directions. This finding correlates with two different strands in Pettit's position on this issue.[10] *On the one hand* (i) it seems as if Pettit wants to use his interpretation of the discursive dilemma as a paradigmatic case and a theoretical resource against the singularist's position. *On the other hand* (ii) the impossibility of an adequate majority aggregation function for judgments or preferences seems to be a problem for the proponent of a non-redundant realism.

Ad (i): We start with outlining the first addressee the discursive dilemma might have. The discursive dilemma attempts to show that there can be cases in votings about proposition-based judgments which are logical connected, in which the voting of a group is not a continuous result of the results of the individual rational voting processes. In Pettit's words: "[t]he 'discursive dilemma' consists in the fact that majority voting on interconnected propositions may lead to inconsistent group judgments even when individual judgments are fully consistent […]." (GA, 46)

The reason for this is that we can choose between premise-based- and conclusion-based-procedures, but they can have reverse results. So there is a difference between the rationality of individuals and the rationality of the group, because they can deviate severely from each other. This seems to be an argument against redundant realism, since redundant realism holds that group-agency talk is not misconceived but readily reducible to individual-level talk (GA, 7). The discursive dilemma now shows that there are cases in which there is no simple relation between individual and group level.

Beyond this, it is not clear at all if the discursive dilemma poses problems for both versions of eliminativism as well, since Pettit defines eliminativism only via a negative thesis, namely that group-level talk is always metaphorical or misconceived, but we do not get to know whether eliminativism takes group-level talk to

[9] Maybe Pettit takes the straw vote procedure in chapter 3.1 to show that his approach provides this feature, but to us it seems that this aims at showing that groups can exhibit reasoning as opposed to mere rationality (cf. GA, 63 ff.). Chapter 3.2. and 3.3 treat the relationship of group members and group agents and therefore presuppose the possibility of group agents.

[10] The first group covers, for the main part, works Pettit published alone (Pettit 2001b, 2004, 2009). In this texts, he seems to vote for option (i). The second group covers works which Pettit published together with Christian List (List and Pettit 2002, 2004, 2011). Here he seems to vote for option (ii).

refer to something that in fact happens on an individual level or whether it takes group-level talk to be meaningless. In the first case, there seems to be no difference to redundant realism, in the second case eliminativism seems to be a very implausible position not many people defend and would therefore appear to be a straw man.

If we follow this route, this immediately leads to a *follow-up question*: What does the discursive dilemma show against the opponents of group-level talk? *Either* it is taken to prove that there are group agents. Group agents, so the thought goes, exist if and only if the premise-based-procedure and the conclusion-based-procedure lead to different outcomes, and if this is the case, then there is no simple reduction from group-level to individual level. This interpretation is backed by Pettit's remarks that "the lack of an easy translation of group-level attitudes into individual-level ones requires us to recognize the existence of group agents in making an inventory of the social world." (GA, 5)

If we follow this line of explicating the role of the discursive dilemma, group agents only exist in those cases in which a discursive dilemma occurs. But this seems strange, because there would not be that many instances of group agency, and the existence of an instance of group agency would depend on the way propositions are interconnected. *Alternatively*, there is the following interpretation: Pettit takes it for granted that group agents exist and the discursive dilemma only makes plausible that they are not in every case just simple aggregations of individual judgments of the members.[11] If this is the case, we still lack a proof for the existence of group agents, because the discursive dilemma is not the candidate for the proof anymore.

In the first interpretation, cases of group agency seem to be very rare, according to the second interpretation the discursive dilemma looses the function to show the ontological assumption, i.e. that group agency is possible or even that group agents exist.[12] But if this is the case, we do not see if Pettit really follows the goal to show that singularism is ontologically wrong and that he wants to convince the eliminativist and the redundant-realist of his position.

Ad (ii): We now come to the second possible addressee of the discursive dilemma. Maybe both, the first and the second interpretation, are wrong and Pettit just wants to give us an internal offer, addressed to those already sympathizing with non-redundant realism and the discursive dilemma is not to be read as a critique of opponents. This leads us back to the second of the two options offered a moment ago. In this case, the discursive dilemma does not pose a problem for singularism, but points at an internal problem for those who want to formulate a coherent version of non-redundant realism. This seems to be Pettit's latest position as it is developed

[11] In the discussion, Pettit explicitly agreed that this is the adequate characterization of his position. But this yields the consequences lined out in the following, namely that Pettit's book is in no way an argument against singularists but only an internal offer for non-redundant realists. Furthermore, this commitment stands in tension with Pettit's contrary suggestion that "the lack of an easy translation of group-level attitudes into individual-level ones *requires us to recognize the existence* of group agents in making an inventory of the social world." (GA, 5, our italics).

[12] This refers back to our problem in identifying what makes up an ontological assumption as opposed to non-ontological assumptions in Pettit.

in *Group Agency*, since the discursive dilemma is generalized to the impossibility theorem, stating that "[t]here exists no aggregation function satisfying universal domain, collective rationality, anonymity, and systematicity" (GA, 50)

These are the four conditions which every aggregation function that serves to derive collective output from individual input has to fulfill. Therefore, Pettit maintains that the conditions for a suitable aggregation function cannot be fulfilled. This reads as an argument against non-redundant realism, since non-redundant realism aims at ensuring the rationality of group agents.

From this, Pettit proceeds without further argument to the claim that "[…] this would be the wrong interpretation of the result. More constructively can be taken to show that, if a group seeks to form intentional attitudes it must relax at least one of the four conditions." (GA, 50)

It is now unclear to whom this statement is directed. If it is only an internal statement towards the fellow non-redundant realists, there seems to be no problem: "If we want to formulate a coherent version of non-redundant realism, we have to relax the conditions for an adequate aggregation function." This implies that if the conditions are not relaxed, there is no way of formulating a coherent version of non-redundant realism. For an internal contribution among non-redundant realists, this is an interesting finding, since it reveals the price for the formulation of a coherent version of this position. But for opponents of non-redundant realism, it seems to be only one further argument to reject non-redundant realism. They do not get to know why they should relax the conditions for an aggregation function Pettit himself introduced.

Since we did not find another part in the book that aims at convincing opponents of non-redundant realism of the necessity to relax these conditions and follow Pettit's way of designing group agents, it seems as if Pettit leaves open the central burden of proof. He does not formulate a positive argument that convinces non-non-redundant realists to giving up there position and sign up to Pettit's. There seems to be no independent proof that "there are group agents" (GA, vii). But as we pointed out at the beginning, it seems as if the burden of proof is distributed equally between singularists and non-redundant realists. So it seems as if Pettit would have to undertake steps in order to refute singularism, which would require more than an internal contribution to non-redundant realism.

9.6 Conclusion

To sum up, we have open questions regarding the foundational work that supports the logical possibility or existence of group agents on broadly two fields. The first field regards the logical space for Pettit's non-redundant realism. Since this space is based on the distinction between methodological and ontological levels, but Pettit's account of methodological individualism seems to be an ontological account, we do not see the logical space for his position. Furthermore, if he binds individualism to contemporary research in psychology, we do not find evidence that his work is backed by contemporary psychology and furthermore we have the impression that his account of rational action is independent of findings in psychology.

The second field regards the proof of the logical possibility of group agents. Here we had problems finding the features of agency developed in the simple model on the level of group agency. It remained unclear whether a suitable aggregation function only ensures rational intentional states or if it includes the third feature, the capacity to process intentional attitudes or whether a group agent does not have representational and motivational states in its own right and a suitable aggregation function takes the intentional states of individuals as input and is itself the capacity to process this, leading to intervention in the environment. In this case the aggregation function would be the analogue to the third feature on the individual level.

Finally, we did not see the function of the discursive dilemma and the aggregation function for supporting the existence of group agents. Either it is an argument against singularism, in which case it seems to leave only rare cases for group agency, or it is only an internal problem among non-redundant realists, in which case Pettit's approach seems to lack an argument against singularism.

References

Gilbert, Margret. 1989. *On social facts*. Princeton: Princeton University Press.

List, Christian, and Philip Pettit. 2002. Aggregating sets of judgments: An impossibility result. *Economics and Philosophy* 18: 89–110.

List, Christian, and Philip Pettit. 2004. Aggregating sets of judgments: Two impossibility results compared. *Synthese* 140: 207–235.

List, Christian, and Pettit, Philip. 2011. *Group agency. The possibility, design, and status of corporate agents.* Oxford: Oxford University Press. (=GA)

Pettit, Philip. 1993. *The common mind: An essay on psychology, society and politics*. New York: Oxford University Press. (=CM)

Pettit, Philip. 2001a. Deliberative democracy and the discursive dilemma. *Philosophical Issues* 11(supp. to Nous): 268–299.

Pettit, Philip. 2001b. Collective intentions. In: *Intention in law and philosophy*, ed. N. Naffine, R. Owens, and J. Williams, 241–254. Dartmouth: Ashgate.

Pettit, Philip. 2004. Groups with mind of their own. In: *Socializing metaphysics*, ed. Frederick Schmitt, 167–193. New York: Rowman and Littlefield.

Pettit, Philip. 2007. Joining the dots. In: *Common minds: Themes from the philosophy of Philip Pettit*, ed. G. Brennan, R. Gooding, and F. Jackson et al., 215–344. Oxford: Clarendon Press. (=JD)

Pettit, Philip. 2009. The reality of group agents. In: *Philosophy of the social sciences: Philosophical theory and scientific practice*, ed. Chris Mantzavinos, 67–91. Cambridge: Cambridge University Press.

Pettit, Philip, and David P. Schweikard. 2006. Joint actions and group agents. *Philosophy of the Social Sciences* 36: 18–39.

Schweikard, David P. 2011. *Der Mythos des Singulären. Eine Untersuchung der Struktur kollektiven Handelns*. Paderborn: Mentis.

Taylor, Charles. 1989. Cross-purposes: The liberal-communitarian debate. In: *Liberalism and the moral life*, ed. N.L. Rosenblum, 159–182. Cambridge: Harvard University Press.

Chapter 10
Pluralism Across Domains

David P. Schweikard

On the face of it, Philip Pettit is committed to an incompatible set of views regarding methodological individualism. In the context of elaborating what he calls the "program architecture," he writes: "We should eschew methodological individualism" (JD, 226). In the context of laying out the fundamental commitments of their conception of group agency, Christian List and Philip Pettit contend and claim it as a merit of the view they term "non-redundant realism" about group agency that it "conforms entirely with methodological individualism." (GA, 4) The two contexts at hand are separable, hence there is no blatant mistake in play here. But the systematically minded reader will still want to know why what appears to be one and the same methodological doctrine is advised against in one context and recommended in another. And if the identity of the doctrines mentioned is only apparent, i.e. if they are in fact discrete, she will want that equivocation to be deleted. This much should be granted to our imagined reader.

Below the surface, however, the statements quoted are not incompatible, but they express methodological commitments (or recommendations) that can be made plausible in their respective contexts. This shall occupy the reconstructive part of this paper. In that respect, I will recast briefly how Pettit understands methodological individualism in those two domains and sketch how these understandings can be kept free of contradicting one another. But Pettit himself gives a cue to seeing how these methodological commitments are connected, and I shall use those remarks to suggest a systematic alignment of the respective methodological doctrines.

Put in preliminary and somewhat rough terms, the systematic options divide into monistic and pluralistic methodological strategies. My suggestion will be, as the title of the paper indicates, to subscribe to pluralism across domains.

The chapter is in four sections. First, I turn to the program architecture as the context in which Pettit advises against methodological individualism (Sect. 10.1).

D.P. Schweikard (✉)
Philosophisches Seminar, Westfälische Wilhelms-Universität, Münster, Germany
e-mail: david.schweikard@uni-muenster.de

S. Derpmann, D.P. Schweikard (eds.), *Philip Pettit: Five Themes from his Work*,
Münster Lectures in Philosophy 1, DOI 10.1007/978-3-319-26103-4_10

In a second step, I take a brief look at the basic structure of the conception of group agency defended by List and Pettit, in which a quite different version of methodological individualism is introduced and adopted (Sect. 10.2). In a third step, I take up a passage from List's and Pettit's discussion of group responsibility in which a connection is drawn between the topics treated in the first two sections (Sect. 10.3), which I will use to suggest a view that is broadly in line with Pettit's different theoretical aims but expressly committed to a pluralist methodology (Sect. 10.4).

10.1 The Program Architecture Contra Methodological Individualism

In his work on causation and causal explanation, Pettit defends the so-called "program model." This account is meant to elucidate not only causation and causal explanation, but also the coexistence and correlation between lower- and higher-level causation as well as the structure of explanations that refer to both lower- and higher-level causes. It forms one branch of Pettit's work on topics more broadly connected with physicalism.

The basic idea is the following[1]: Causes produce effects in virtue of some of their properties. These are the properties typically mentioned in plausible causal explanations, i.e. statements that specify in virtue of which of its properties one event (including the properties of the entities involved) brings or brought about another. Such causally relevant properties, Pettit argues, "may be nested in relation to one another, with more general properties figuring at the higher end and more specific ones at the lower." (JD, 220) It is the notion of causal relevance in play here that prompts the introduction of the notion of programming:

> A property that is causally relevant to the appearance of a certain type of effect will program for that production in the following sense. It will be capable of being realized across a range of variations in the realizing event: this, as the bolt of lightning may be realized across variations in the magnitude of the electrical discharge. No matter how it is realized, however, the realizing event will produce or help to produce—or be likely to produce or help to produce—the type of effect at issue: this, in the way that the different electrical discharges that may constitute a bolt of lightning will each tend to produce the electrical failure. And, finally, it is a realizer of that property, a particular bolt of lightning, with a particular magnitude, which actually does produce or help to produce the effect. (JD, 220)

Programming occurs both on the type-level, where in the case of a causal event's producing a certain type of effect "a causally relevant property programs for the production of that type of effect," and on the token-level, where the instantiation of a certain property "programs for the appearance of that type of event." (Ibid.) On the program model, such property-instances are causally efficacious, whereas properties qua types have causal relevance.

[1] This reconstruction is based on JD, section I.

A specific problem the program model is designed to deal with is posed by the coexistence of higher-level and lower-level causes. How can there be, the physicalist may ask, any other kind of causality beyond that occurring between microphysical events? Could there be macrophysical, psychological, or sociological causality that deserves to be recognized as causality at all? And if such higher-level causality exists, how does it relate to the causality at the lower level(s)?

The key to making room for forms of higher-level causality is an account of "how properties at different levels can be causally relevant." (Ibid.) The distinction between levels, and properties at different levels, is cashed out in terms of supervenience: "One property is higher in level than another if the realization of properties of that type is superveniently determined by the realization of properties of the other type." (Ibid.) According to the model of causality under discussion, this is transferred to causally relevant properties at different levels.

> Whenever a property programs for the production of an event, it will display higher-level causality so far as there is a lower-level property that realizes the higher level, in the usual pattern of supervenience, and that lower-level property programs at a more specific level for the production. (JD, 221)

In this sort of situation, the higher-level property and the lower-level property are both causally relevant in that they program for the production of the effect at their respective level of specificity. The hierarchy described in these terms is one in which programming occurs on different levels, but without conflict. Pettit elaborates that whereas it may be tempting to say that the most specific and in that sense most fundamental level is also ultimate in the sense of deserving to be prioritized in the metaphysics of causality, "there need be nothing special or different about such maximally specific programmers, apart from the fact that they come at the bottom of the program architecture." (JD, 222)

Skipping the reconstruction of further details of the account, especially the idea of co-programming without co-instantiation (JD, 222–4), let us consider the view of causal explanation Pettit presents in the same context. One decisive step in this regard consists in treating each description of a process in which programming properties are identified as providing information about an event's causal history, i.e. as a variety of causal explanation (JD, 225). As there can be programming properties on different levels, there can be causal explanations on different levels, each on a definite level of specificity, providing the respective form and body of information. These bodies of information – e.g. the information that a bolt of lightning caused the fire, and the information that a bolt of lightning of a specific magnitude *M* caused the fire – serve different purposes, and Pettit takes this fact to imply "that there is no single right way of explaining an event. The program architecture allows for explanation at different levels, and each explanation will have its own merits, providing information that answers distinct questions." (JD, 226)

In line with the repeated advice against prioritizing lower-level or more basic programming properties, Pettit advises to "eschew explanatory reductionism, which would favor going to finer and finer grains of information in seeking the explanation of any event." Instead, Pettit continues,

> [w]e should embrace what I call ‘explanatory ecumenism’ or pluralism. This is an important lesson in many areas, particularly in social science. There is no reason to have to choose between favoring individual-level accounts, for example, and explanations of a more structural, higher-level kind: explanations such as that which invokes urbanization to explain secularization, or a rise in unemployment to explain an increase in crime […]. We should eschew methodological individualism, even if we embrace [.] ontological individualism […]. (JD, 226)

This passage contains the statement targeted with this reconstruction of Pettit’s account of causation and causal explanation. We can understand it as a recommendation against reductionist methodological strategies that treat only explanations at the most basic level as informative. In so far as this view implies treating precisely one level as the adequate reference for an informative causal explanation, we can take it to be suggesting a monistic methodology. Against this Pettit recommends embracing explanatory ecumenism or pluralism. Adapted for the variety of explanations aimed at in the social sciences, the advice goes against methodological individualism, which could be construed as a reductionist and monistic methodology. However, just as the program model of causation is compatible with physicalism, this rejection of methodological individualism is compatible with ontological individualism. The message here is that a basic, monistic ontological commitment – e.g. to physicalism or to individualism – does not imply a commitment to a monistic methodology.[2]

The explanatory strategy thus suggested is ecumenical and pluralistic not just in that it allows for a plurality of causal explanations, it is also fit to incorporate non-causal explanations. On Pettit’s account, giving an explanation requires invoking a controller and there are important kinds of non-causal control. If a causal process leads to a certain sort of effect, Pettit explains, “we may say that it controls for that effect” (JD, 226); but beyond the typical case of active control, there is also a kind of controlling influence that is distinct from active causal influence. Pettit calls such controllers ‘virtual controllers,’ where ‘virtual’ means that the controller in question has a standby role in a causal process and only steps in if the normal controller does not play its causal role. For instance, deep seated convictions may guide much of what an agent does in this sense of virtual control if they are not made explicit or regularly reflected upon in case-by-case decisions, but intervened if conscious deliberation ran counter to them.

[2] Of course, much more needs to be said about this combination of monistic ontological and non-monistic methodological commitments. The reflections offered here focus on the plausibility of methodological pluralism, remaining silent on the question whether the commonly held monistic ontology is without alternatives (see Turner 2010).

10.2 Group Agency Cum Methodological Individualism

I now turn to List's and Pettit's account of group agency and, in particular, to the invocation of methodological individualism as a fundamental commitment. Indeed, at times List and Pettit suggest to treat adherence to methodological individualism as a condition of adequacy for a convincing account of group agency. That decision may be contextually justified, but what are we to make of the apparent tension in which it stands to the rejection of methodological individualism we looked at in the previous section?

Here is my short answer to this question: The methodological strategy suggested in the context of the program architecture is also recommended for and indeed operative in the domain of group agency. The view behind the label 'methodological individualism' as it occurs in the exposition of List's and Pettit's account of group agency is not only or primarily a view about explanatory strategies or about methods of explanation in social science, it is an ontological commitment. As such, the view is compatible with pluralism about explanation, which is precisely what List's and Pettit's own invocation of the program model and their use of the concept of virtual control suggests. – So much for the rough and ready, preliminary version of the argument, which the remainder of the contribution is meant to spell out.

Let us begin by taking a brief look at List's and Pettit's account of group agency. The account is extremely rich, its exposition and defense are to be praised for their stringency and precision, so it comes as no surprise that I can only give a very selective reconstruction here and have to set aside all references to other accounts of group agency.[3] In doing so, I shall focus on how List and Pettit introduce the thrust of their account and their definition of and commitment to methodological individualism.

On the most general level, List's and Pettit's project (in *Group Agency*, here referred to as 'GA') is to develop "a theory of group agency, and [to explore] its implication for the organizational design of corporate entities and for the normative status they ought to be accorded." (GA, 2) In opposition to views that deem all reference to group agents as metaphorical and that require that group agents be eliminated from theorizing in social science and social philosophy (including the philosophy of action), List and Pettit defend a non-eliminativist or non-singularist position and argue in favor of a specific variant of realism about group agency. This view, however, shares an important basic commitment with eliminativist and singularist positions: the commitment to methodological individualism. List and Pettit understand it as

> the view that good explanations of social phenomena should not postulate any social forces other than those that derive from the agency of individuals: that is, from their psychologically explicable responses to one another and to their natural and social environment. (GA, 3)

[3] For this, see Schweikard (2011, part III.).

Dropping the qualification 'methodological,' and drawing a parallel to the role of physicalism in biology and psychology, they go on to say that "individualism says that economic and social explanations should resist any appeal to psychologically mysterious social forces." (Ibid.) Both these specifications provide a negative guideline for adequate (or 'good') explanations in stating that they should *not* postulate or appeal to social forces that are either detached from individual agency or psychologically mysterious.

Now, the decisive move at that stage is to detach the commitment to methodological individualism, so defined, from that to eliminativism. List and Pettit do this by arguing that whereas the latter implies the former, the reverse does not hold, so that there is room for a view that subscribes to methodological individualism and is non-eliminativist in acknowledging that there really are group agents that have significance in social life and reference to whom can be justified. They take this view to be in line with methodological individualism, for it does

> not introduce any psychologically mysterious forces. As the agency of individual human beings depends wholly on the configuration and functioning of biological subsystems, so the agency of group agents depends wholly on the organization and behavior of individual members. Despite being non-eliminativist, this picture conforms entirely with methodological individualism. (GA, 4)

List and Pettit argue that (some) groups deserve to be accorded significance and a very specific form of agential autonomy, yet this realist conception does without "compromising the individualist claim that no psychologically mysterious forces should be invoked in giving an account of the social world." (GA, 6) The argument of the first part of their book proceeds by showing how (some) groups can exhibit the features that warrant their recognition as agents in their own right. While much more could be said about that argument, let me note only in passing that it is not clear whether there is any real opposition to the rejection of 'psychologically mysterious forces.' At least to my knowledge, no philosopher – especially no-one writing in the twenty-first century – seems to be ready to subscribe to that sort of position, neither in view of its ontological claim nor in view of the methodological prescription aligned with it.

10.3 Group Responsibility and Higher-Level Control

Against the background of this partial reconstruction of List's and Pettit's taxonomy, I now turn to one specific argument, presented in the seventh chapter of *Group Agency*, which highlights the particular normative status of groups that qualify as agents. The context for that argument is the question as to whether and, if so, under what conditions group agents can be suitable targets of ascriptions of responsibility. Questions of collective and corporate responsibility have been debated in social philosophy at least since World War II, and indeed, they are both practically pressing and provide important applications for theories of group agency.[4] List's and Pettit's account can be read as approaching these issues within an action theoretic framework that is comparatively much more precise that those presupposed by most other accounts of collective and corporate responsibility. As with groups' potential status as agents, List and Pettit work with a list of requirements candidate groups have to fulfill in order to count as 'fit to be held responsible' (GA, 115-7). I shall first reproduce this list and then focus on their explanation concerning how groups can fulfill the third of these requirements.

The three requirements a group needs to fulfill in order to be fit to be held responsible are the following:

> First requirement. The group agent faces a normatively significant choice, involving the possibility of doing something good or bad, right or wrong.
>
> Second requirement. The group agent has the understanding and access to evidence required for making normative judgments about the options.
>
> Third requirement. The group agent has the control required for choosing between the options. (GA, 158)

Let us accept the result of List's and Pettit's discussion of the first two requirements for the time being and assume that they can be met by quite ordinary group agents. The third requirement, which we can also call the control requirement, poses a problem we are already familiar with as it turns on the question as to how the group agent (as such) can be ascribed the relevant control over the choice between options and the respective performance. How can the group agent be said to control actions that are performed by its members? Clearly, List and Pettit do not want to claim that an individual who performs a certain action qua member of a group agent, i.e. in the name of the group, does not control what she does. Since all actions by group agents are performed by individuals, some of them as joint actions, the challenge is to say how they can be subject both to the control of the group agent and to the control of the individual agent who performs the action (GA, 160-1).

This is, as List and Pettit expound, an instance of the more general problem of multi-level causality. In this context, they use a case in which water in a closed flask is brought to the boil, which brings about the breaking of the flask (GA, 161). As the water boils, so a simplified account of the case goes, some molecule is fast enough at the right place to bring about a crack in the glass when it hits the surface. So what we have here as relevant factors in the story of the collapse of the flask is the water temperature and the molecule that brings about the (first) crack. Both are causally relevant and it should not surprise us that List and Pettit invoke the program model to give an account of their relationship:

> The higher-level event – the water being at boiling point – 'programs' for the collapse of the flask, and the lower-level event 'implements' that program by actually producing the break. The facts involved, described more prosaically, are these. First, the higher-level event may be realized in many different ways, with the number, positions, and momenta of the constituent molecules varying within the constraint of maintaining such and such a mean level of motion. Second, no matter how the higher-level event is realized – no matter how the

[4] See, for instance, May and Hoffman (1991).

> relevant molecules and motion are distributed – it is almost certain to involve a molecule that has a position and momentum sufficient to break the flask. And, third, the way it is actually realized does have a molecule active in that role.
>
> Given the fulfillment of these conditions, we can say that the water's being at boiling temperature 'programs' for the breaking of the flask, whereas the molecule's behaving as it does 'implements' that program, playing the immediate productive role. Both programming and implementing are ways, intuitively, of being causally relevant and so it makes sense, depending on context, to invoke one or the other in causal explanation of the effect. Information about either antecedent, higher-level or lower-level, is significant for the causal history of the event [...]. (GA, 162)

List and Pettit take this to provide a cue for dealing with the problem of a group agent's control over the actions individuals perform in its name. The solution they propose to this problem has it that the group agent "can share in that control so far as it relates as a 'programming cause' to the 'implementing cause' represented by the enacting individual." (Ibid.) And they continue as follows:

> The temperature of the water controls for the breaking of the flask so far as it ensures, more or less, that there will be some molecule, maybe this, maybe that, which has a momentum and position sufficient to trigger the breaking; the molecule itself controls for the breaking so far as it ensures that this particular crack materializes in the surface of the flask. *Things may be perfectly analogous in the case of the group agent.* The group may control for the performance of a certain action by some members, maybe these, maybe those. It does this by maintaining procedures for the formation and enactment of its attitudes, arranging things so that some individuals are identified as the agents to perform a required task and others are identified as possible back-ups. Consistently with this group-level control, those who enact the required performance also control for what is done; after all, it is they and not others who actually carry it out. (GA, 162–3, my emphasis)

This concludes the argument in support of the claim that group agents can meet all three requirements for fitness to be held responsible. The view that takes shape in this passage and is elaborated further in the seventh chapter of *Group Agency* is particularly attractive, in that it provides a way of dealing with the group's and the enactors' respective responsibilities that is grounded in the metaphysics of group agency. This view neither in principle relieves groups (or corporations) of their responsibility for what is done in their name, according to their statutes and agendas, nor does it let individuals qua members off the hook all that easily by saying that the responsibilities rest only with the group. So just as List's and Pettit's account of group agency sits between eliminativism and emergentism, their account of group responsibility sits between strong versions of individualist and collectivist accounts.

But what are we to make of the 'perfect analogy' between the water temperature's controlling for the breaking of the flask on the one hand and the group agent's controlling for the performance of a certain (individual or joint) action? In particular, what methodological lesson are we to learn from the analogy?

10.4 For a Methodological Pluralism

Let us assume that there is a point to the analogy, so that it helps us gain actual insight into the structure of group agency and group agents' virtual control over the actions. As I indicated, the view of group responsibility that emerges is very appealing.

But the question primed by the considerations in this paper regard the methodological commitment in play. In particular, the question is why we should stick with methodological individualism if we find it persuasive to treat the group's virtual control just as we treat the temperature's virtual control in the case of the breaking flask, i.e. as an instance of programming. Since it is precisely an account of *multi*-level causality, the program model not only speaks against reduction, it requires to take explanations that refer to different levels but are all the same informative equally seriously, which implies that higher-level explanations are not to be jettisoned in favor of lower-level explanation.

Now, in *Joining the Dots*, Pettit explains that if we accept the program model we should 'eschew methodological individualism' and instead adopt ecumenical or pluralist explanatory strategies. I want to suggest nothing more and nothing less than this: in virtue of the analogy between programming and implementing in cases such as those of the bolt of lightning or the breaking of the flask on the one hand, and programming organizational structure and implementing individual action by members in the case of group agents, we are well-advised to be *pluralists across domains*.

Such pluralism would not amount to postulating anything like psychologically mysterious forces, thus it would in that sense contradict the central aspects of List's and Pettit's definition of methodological individualism. But it would certainly make for an explanatory apparatus that is fit to deal with the complexities of group agency, and probably social phenomena more generally. Furthermore, adopting methodological pluralism in the domain of group agency would calm the systematically minded reader of Pettit's work who is worried by the equivocation we cited at the outset.

References

List, C., and P. Pettit. 2011. *Group agency – The possibility, design, and status of corporate agents*. Oxford: OUP (=GA).

May, L., and S. Hoffman (eds.). 1991. *Collective responsibility – Five decades of debate in theoretical and applied ethics*. Savage: Rowman & Littlefield.

Pettit, P. 2007. Joining the dots. In *Common minds – Themes from the philosophy of Philip Pettit*, ed. G. Brennan, R. Goodin, F. Jackson, and M. Smith. Oxford: Clarendon Press (=JD).

Schweikard, D. P. 2011. *Der Mythos des Singulären – Eine Untersuchung der Struktur kollektiven Handelns* [The myth of the singular – A study of the structure of collective action]. Paderborn: Mentis.

Turner, J. 2010. Ontological pluralism. *The Journal of Philosophy* 107(1): 5–34.

Chapter 11
Which Liberalism, Which Republicanism? Constructing Traditions of Political Thought with Philip Pettit

Sven Lüders and Johannes W. Müller-Salo

11.1 Introduction

Throughout his books on political theory, Philip Pettit refers to certain elements of the republican tradition to support his own arguments. The first chapters of *Republicanism* and *On the People's Term* present a grand historical narrative[1] and name writers important for republican theory building (R, 19f; OPT 5–8). In the further course of these books Pettit repeatedly mentions these theorists as forefathers of his own republican theory. Thereby he maps the history of political thought. According to him, republicanism can be separated from two other important traditions, "liberalism" and "populism" (R, 7–11; OPT, 8–18). Pettit's main concern is the rejecting of liberalism; he scarcely mentions populism at all. Since Hobbes liberalism has gained widespread support and displaced republicanism in the nineteenth century (R, 37 f.). Confronting this process, Pettit argues for the need of revitalizing republican thinking.

In this paper we examine Pettit's historical account. To do so, we will look at the two groups of theories he distinguishes from republicanism. We claim that the way he uses the history of political thought is questionable and not supportive for his own theory. According to him, the links to the tradition should not be understood as some kind of argument from authority, but as an attempt to reinforce his arguments (R, 10; OPT, 18 f.). Using tradition like that can be adjuvant, but only under certain

[1] In this text, our main focus will be on *Republicanism*. In this earlier book Pettit gives the most detailed historical account and we cannot see that he has changed this account ever since.

S. Lüders (✉)
Philosophisches Seminar, Westfälische Wilhelms-Universität, Münster, Germany
e-mail: sven_lueders@yahoo.de

J.W. Müller-Salo
Kolleg-Forschergruppe Normenbegründung in Medizinethik und Biopolitik,
Westfälische Wilhelms-Universität, Münster, Germany
e-mail: j.mueller-salo@uni-muenster.de

S. Derpmann, D.P. Schweikard (eds.), *Philip Pettit: Five Themes from his Work*,
Münster Lectures in Philosophy 1, DOI 10.1007/978-3-319-26103-4_11

conditions: Namely if the presented narrative is a widely accepted one, which can be used without raising further questions concerning the appropriateness of this historical interpretation. Methodologically speaking such narratives provide orientation and help the reader in forming expectations. Naturally there does not exist *one* valid single interpretation of theorists like Hobbes. Nevertheless, there are interpretations that are more typical and widely shared than others. We think that at some points Pettit's reconstruction of tradition is not of a typical kind, but rather exceptional. In these cases we think Pettit should either explain why he presents the history of philosophy the way he does, or better: leave all historical connections out. This is especially important if his historical interpretation can be used against his systematic arguments, as we will try to show below. In other words, we ask for the criteria that have been used in building the three mentioned groups of theories. We follow this main question through four sub-questions: two in the field of liberalism and another two with regard to the republican tradition.

11.2 Which Liberalism?

Liberalism, as Pettit himself underlines, is "a broad church" (R, 9; 50). It is a label used for many theories that aim at different conceptions of polity and society. To avoid misunderstandings we accept Pettit's description: "I think of liberals as those who embrace freedom as non-interference" (R, 9). In contrast, Pettit describes his own republican concept of freedom as non-domination. On his interpretation persons will be free in the sense of liberalism, if and only if they are free of any physical coercion as well as any threat of such coercion. It is not important who the person or institution is that uses such force. This is the main difference between liberalism and republicanism, as Pettit tries to explain using the traditional republican example of master and slave: Let us suppose that A is the slave of a benevolent master B, who does not interfere with A's options of action. This case will be one of domination, but none of interference. There exists a person B, who has the power of interfering arbitrarily in A's life, therefore the republican criterion for freedom is not fulfilled: "I suffer domination to the extent that I have a master; I enjoy non-interference to the extent that that master fails to interfere" (R, 36). While this is a case of *domination* without *interference*, the opposite of a situation of *interference* without *domination* is imaginable as well. It will be a case in which someone interferes with the actions of another person under the republican condition "that the interference promises to further my interests, and promises to do so according to opinions of a kind that I share" (R, 23). The interference caused by a well-ordered republican state is of such non-dominating kind. Pettit's line of argument seems to be clear: With the liberal conception of freedom in mind, it is not possible to criticize social practices like slavery or the unequal treatment of women and men; furthermore liberals have to judge any kind of state organization as compromising freedom. Building on this interpretation of liberalism we want to mention two points that prevent us from agreeing with Pettit's view.

11.2.1 Conditions for Realizing Freedom

We think that the fundamental difference between the two conceptions of liberty presented above is an abstract difference that points at two different ideals of freedom. The central question is: What happens if we move from this description of ideals to problems of institutional design and conditions of realization? Pettit himself mentions the starting point for any liberal theory: "To advance freedom as non-interference will be to remove interference as far as possible, and to expand as far as possible the sphere of uninterfered-with choice" (R, 83). It is important to emphasize the phrase 'as far as possible': All human beings live in societies; therefore other persons will necessarily interfere with them. Thus the liberal problem is to identify the kind of interference that is the least evil.

Two possibilities present themselves: Either the continued interference of state institutions is seen as the greater threat to liberty than the arbitrary interference of other people, or the answer is contrariwise and the possibility of other people interfering arbitrarily seems to be worse than the interference caused by official legislation. The first position is quite distinguishable from republicanism and hails some form of liberal anarchy, but unfortunately there are no liberals who embrace that position. The liberals since Hobbes took the second route, making use of the idea of a fictive state of nature illustrating that leaving such a state and forming continuously interfering public institutions is always the better option. Pettit is right in claiming that in this view law seems to be a restriction on liberty,[2] but nevertheless liberals have ever since argued forcefully for the foundation of such a freedom-restricting institutional setting.

If this line of interpretation is correct, then the differences in the republican and liberal conceptions of freedom under the condition of statehood become dubious. This will be clear, if we return to the example of the slave who enjoys freedom, because his benevolent master decided not to interfere with his actions (R, 23). Thinking about institutional design, the liberal wants to create a society which allows for as little interference as possible. Given this intention, how probable is it then that he will accept such an institution as slavery? It is true, there might exist a benevolent master who will never interfere with the options of his slave, but the belief in such virtuous men will never be that strong that the liberal will be ready to accept slavery as a social institution. This famous example is misleading. Pettit writes at the end of *Republicanism*: "Freedom as non-interference was weak on the social front, because it could not be used to criticize the contemporary relationship of master to servant or of husband and wife; however powerless they were, servant or wife would be free in this sense provided the master or husband stayed his hand"

[2] See e.g. R, 50, and the differences between liberals and republicans concerning the understanding of law, R, 84: "The difference between the two ideals on this front connects with their different views of law. Devotees of freedom as non-interference view legal or state coercion, no matter how well bounded and controlled, as a form of coercion that is just as bad in itself as coercion from other quarters; if it is to be justified, then that can only be because its presence makes for a lesser degree of coercion overall."

(R, 298). In our view, this and other similar statements can only be interpreted in one way: If liberals are to criticize these unfair social structures like slavery – and obviously, as Pettit himself admits, they *do* criticize these structures – they can only do that because they are committed to other values besides freedom as non-interference. These values can be e.g. justice, fairness and equality. To us, this interpretation of the liberal core idea seems misleading. Nevertheless, this is the only story the example of master and slave can tell. Pettit is surely right in claiming that early liberals did not criticize unfair social structures – but the same is also true for the republican tradition: Cicero and Jefferson were slaveholders themselves. Pettit mentions several times that the republican ideal was traditionally an ideal for rich white men, but claims rightly that it is possible to develop this ideal further into an egalitarian conception that includes every human being. Nevertheless it is incomprehensible why the same should not be true for liberalism as well. Modern liberalism could take social issues seriously, although the 'fathers of liberalism' (such as Bentham) oversaw this social dimension.

Our main question has to be restated here: If liberals ever since agreed upon the necessity of a state that protects individual rights – rights that are obviously violated under conditions of slavery and male dominion – then it remains dubious why Pettit dissociates himself from the liberal heritage. Does he make the mistake of equating 'liberalism' with some kind of 'radical Manchester liberalism' that was influential during the time of capitalist expansion and industrialization?

11.2.2 Liberal Traditions? Hobbes, Filmer and Locke

Secondly we think that Pettit's construction of the history of liberalism is questionable. At various stages he describes liberal positions as ideologically susceptible. Let us for example take a look at the following sentence:

> The notion of freedom as non-interference first became prominent, so I suggest, in the writings of a group of thinkers who had an interest, like Hobbes and Filmer, in arguing that all law is an imposition and that there is nothing sacred from the point of view of liberty about republican, or even non-authoritarian, government. The group I have in mind were all opposed to the cause of American independence (R, 41 f.).

Pettit does not claim that this group of thinkers had certain 'arguments', but 'interests'. In mentioning the American Revolution he creates the impression that liberalism is useful in supporting positions whose fallaciousness has been shown by history itself; that liberalism can be used and was actually used in suppressive structures like the British colonial system in America as a legitimation of unjust rule. We think it is better to avoid such assumptions because they are not helpful in arguing for or against a philosophical position but attempts to persuade someone of the intuitive plausibility or implausibility of a theory.

Of special interest in Pettit's construction of the history of liberal thought is his treatment of Hobbes and Filmer. According to him, Filmer has been the only thinker that hailed Hobbes' conception of freedom as non-interference in the early modern times (R, 38; 41). Pettit rightly claims that some sentences in Filmer's work can be read as an embrace of the concept of freedom as non-interference. Nevertheless, the main ideas Hobbes and Filmer are about to defend are of a totally different kind. The question of state legitimacy is at the center of Hobbes' political thought.[3] Hobbes' answer is a new, radical one: The voluntarily and announced agreement of individuals is the only true fundament of state authority and legitimacy. Liberals have ever since built upon this idea while avoiding the consequences of the despotic state of *Leviathan*. Exactly this idea Filmer tried to reject arguing in favor of a divine fundament of state authority. God has given the earth and every creature that lives on earth to Adam,[4] who became the first absolute monarch on earth. This absolute power was passed on as an inheritance from generation to generation. The actual holder of this power – received through the right of primogeniture – is the Stuart king of England.[5] It is obvious that the main aims Hobbes and Filmer are arguing for are totally different from one another. As said above, there are always different possibilities of interpreting classical texts. However, given the differences between Hobbes and Filmer, Pettit's position is unconvincing as long as he cannot provide a convincing argument in favor of his interpretation. We do not know of any liberal who would include Filmer in a history of his own position. To be sure, Pettit never labels Hobbes and Filmer as 'liberals' themselves, but if Hobbes and Filmer were the first to make use of the concept of freedom as non-interference and if this concept is the core idea of liberalism, the liberal history would have to start with them.

Pettit not only includes some authors into the liberal tradition, he excludes others as well. For example, Locke is sometimes considered as a representative of republicanism: "John Locke is a good representative of the commonwealth tradition, though the originality of his rational, contractarian perspective gives him a special status" (R, 40). If Pettit mentions this, he should mention as well that this 'contractarian perspective' is Hobbes' inheritance.[6] Liberals normally perceive Locke as one of their most important ancestors, Crawford MacPherson even claimed in his provocative and controversial book *The Political Theory of Possessive Individualism*

[3] As Pettit (OPT, 141f) himself emphasizes.

[4] Filmer bases his theory on an interpretation of some biblical verses, especially Gen 1:28.

[5] "It is a truth undeniable, that there cannot be any multitude of men whatsoever, either great or small, though gathered together from several corners and remotest regions of the world, but that same multitude, considered by itself, there is one man amongst them that in nature hath a right to be the King of all the rest, as being the next heir to Adam, and all the others subject unto him." (Filmer 1949, *Anarchy of a Limited or Mixed Monarchy*, 288).

[6] Locke certainly used some of Hobbes'arguments without accepting the whole Hobbesian conception. Therefore, Ashcraft 1987, 41 could mention Locke's "certain fascination with Hobbes' philosophy." Laslett has given a convincing analysis of the relationship between Hobbes and Locke (Laslett 1988, 67–92). He emphasizes that Locke beyond doubt had "a great deal in common" with Hobbes.

that Locke is a defender of radical, early modern capitalism.[7] On the other hand Pettit claims that the republican tradition is associated even with "'old liberals' such as John Locke" (Pettit 1999, 166). Is Locke now a republican or a liberal in Pettit's view? His treating of Filmer and Locke exemplifies our main difficulty: The rather untypical lines of tradition he draws blur the theoretical differences between republicanism and liberalism Pettit has developed before.

11.3 Which Republicanism?

What is more, Pettit distinguishes his theory of republicanism from other conceptions which are usually labeled as 'republicanism'. Hereby he means in particular communitarian positions that he calls 'populist' (R, 7–8; OPT, 8–9.).[8] The alternatives held by such positions are either ignored by Pettit (e.g. OPT, 5–6) or put aside with reference to their incompatibility with the core ideas of (his) republicanism (OPT, 15).

In two areas it can be seen where Pettit draws the line: the scope of his conception and the participation it allows.[9]

[7] Macpherson 1962, see e.g. 221: Locke "has erased the moral disability with which unlimited capitalist appropriation had hitherto been handicapped. […] But he does even more. He also justifies, as natural, a class differential in rights and in rationality, and by doing so provides a positive moral basis for capitalist society."

[8] Pettit makes the assumption that his communitarian opposition from Jean-Jacques Rousseau over Hannah Arendt to Michael Sandel shares a common tradition (OPT, 11–18, esp. 12). Unfortunately we cannot analyze this assumption in detail. But we must doubt, if it is not anachronistic to call Rousseau a communitarian. Also this interpretation neglects the strong (systematic) differences between the individual authors. One example is Pettit's handling of Arendt. She surely does not have a wish for a "close, homogenous society" (R, 8). On the contrary she was deeply horrified of the 'Volksgemeinschaft' of the Nazis, which was an attempt to realize this wish, and analyzed and criticized it in 'The Origins of Totalitarism' (Arendt 1951).

[9] These comments are loosely based on Iseult Honohan's introduction into republicanism (Honohan 2002). In her book she analyses four core themes of the republican tradition of political thought, whose increased presence she attributes respectively to four eras of the history of the republican tradition: 1. virtue, 2. freedom, 3. participation, 4. recognition. In each era she also asks which scope a conception has, meaning who it addresses, who is considered to be a citizen and how many people participate in the corresponding republic.

Pettit undoubtedly recognizes and reproduces the theme of freedom, but it is not entirely clear how he positions himself to the other themes of the tradition. Due to the shortage of space we limit ourselves to the questions of scope and participation. Nevertheless we think that Pettit's conception has equal problems with the theme of virtue.

11.3.1 Scope

The first evidence for the difference between Pettit's conception and the so called 'populist' line of the republican tradition is their difference in scope. Aside from his references to the particular context of the pre-modern examples (e.g. R, 134), Pettit does not state whom his theory addresses. Examples of states, which are likely to realize a republican policy conforming to his ideas, are also missing.[10] This is not in line with the habits of the republican tradition, as e.g. Iseult Honohan presents them. According to her, the republican authors did not present a general political theory. Instead they addressed a particular society (Honohan 2002, 6–7), especially Cicero's Rome and Machiavelli's Florence. The problem is not the universalization of republicanism – there may or may not be good arguments for that. The problem lies in the unwillingness to make such an argument. Why should it be the aim of republicanism to be a political theory for every human being and how is this possible? We are not convinced of Pettit's attempt to move from the particular background of the republican tradition to a universalist interpretation.[11]

This gap is especially important in Pettit's argument with his communitarian opponents. He accuses them that their focus on contextualizing political theory in a particular (e.g. western) society is not compatible with the "contemporary, pluralistic society" (R, 96). Pettit believes in contrast to his communitarian critics that his ideal of freedom as non-domination is simultaneously neutral and motivating. He accuses his critics of idealizing a homogeneous society.[12] However, this accusation is not imperative. Granted, particular ideals and values homogenize society, but not necessarily in such a measure that they would be incompatible to modern societies. Pettit's accusation is also beside the point of his communitarian critics. Pettit simply avoids their argument of collective ideals needing to be based in a collective identity by talking about the pluralistic society. Instead, Pettit is convinced that he can eliminate the doubts of communitarians over the course of his argument (R, 97). He thinks that his deliberations on the republican community (R, chapter 4) and on virtue (R, chapter 8) in particular are suitable to show that the ideal of freedom as non-domination is sufficiently motivating.

One important theme in the persuasion of the communitarians is the proposition that freedom as non-domination in fact already is a communitarian ideal (R, 120–

[10] There may be the exception of Pettit's book covering the reception of his philosophy in Zapatero's Spain (Martí and Pettit 2010). However, Pettit's republicanism appears in the book as a general theory which was applied in Spain, not as a theory for Spain (or another western country).

[11] We do not think that the gap between Pettit's theory and the classical texts can easily be bridged with an adoption of a third way between particularism and universalism. Such a solution would suggest for example that freedom as non-domination is a universalistic goal only insofar as it is implemented into a particular society. But the problem of the perspective would remain: Whereas the classical authors addressed their societies, Pettit would still aim for neutral point above a concrete society.

[12] Pettit's republicanism is compatible with pluralistic societies because of the adoption of egalitarian ideas. From his point of view it can be said that his opponents (although he doesn't call them republicans) remain in an older, pre-modern form of republicanism (R, 96–97).

121). Republican freedom, so he claims, is a common good, because all citizens belong to at least one group that is in danger of losing its freedom (R, 124). Secondly, freedom as non-domination is not an atomistic but an inter-subjective ideal because it needs to be shared by the group to which the individual citizen belongs (R, 125). Pettit goes even further by stating that the motto 'liberté, égalité et fraternité' attributed to the French Revolution counts as an expression of the republicanism he prefers (ibid.). In addition to the connection between freedom and egalitarianism, which distinguishes his republicanism from pre-modern theories, the motto expresses also a connection between freedom, egalitarianism and the idea of a republican community.

A communitarian will nonetheless find it difficult to accept freedom as non-domination as a political ideal. The statement above referring to the French Revolution stands particularly in contrast with Pettit's preoccupation of giving a definition of freedom.[13] Pettit himself admits the latter statement (OPT, 7 f.). The missing (explicit) link of Pettit's conception to particular contexts nevertheless leaves the communitarian also in massive doubts. Usually those contexts constituting a community are of a particular kind like religion and nation.

To return to our main question: If republican thinkers ever since acknowledged the necessity of taking into consideration local circumstances of that one republican system they were writing for, how can Pettit ignore this important line of thought without further explication and call his position nevertheless republican?

11.3.2 Participation

The most important boundary of Pettit's philosophy in relation to other positions that understand themselves as republican is their central postulation of participation (R, 8). Even if Pettit himself acknowledges that participation is an important component of republican freedom (OPT, 22), he prefers the idea of contestability to that of participation. Contestability should ensure two things: (1) the identification of the citizen with democratic decisions and (2) the legitimacy of the democratic process (R, 184).[14] Following Pettit's own conception, participation is not an end in itself and not an expression of positive liberty, but rather an instrument for the perpetuation of freedom as non-domination. Pettit is in particular skeptical of the postulation

[13] Honohan's portrayal of Pettit: Honohan (2002, 8, 59, 183–187).

[14] At this point it must be asked, if the underlying theory of democracy is not contra-intuitive. In this theory democracy is more about the selection of laws that can withstand the contestations of the citizens than about the making of new laws (R, 201–202). This contrasts with the reality of many processes of democratic decision making, which are sometimes more about the identity of a political community. If we follow Pettit's theory these are devaluated to mere symbolic gestures, because the consequences of them are often unclear. But in fact they are an integral part of the political system. If one accepts the idea of participation, one does not run into these problems necessarily because the idea of participation helps in bringing out the necessity for the citizens to identify with democratic decisions.

of a direct democracy because he sees in it the danger of a "tyranny of a majority" (ibid.).[15]

Pettit suspects that participation is an ideal deriving from 'populistic' ideas of positive freedom (R, 81). He uses the concept of participation on the background of Rousseau's political philosophy and sees it inevitably linked with the 'volonté générale'. Consequently he arrives at the conclusion that the ideal is not up to date. This conclusion misses an assessment of the link between the evaluation of Rousseau's philosophy and the ideal of participation. Newer positions, in particular Benjamin Barber (2009), but also the otherwise praised (OPT, 12) Jürgen Habermas (1998, 129–133), are not discussed. There is also no discussion of ideas that do not derive directly from Rousseau's 'volonté générale' like Arendt's reception of Jefferson's proposal of a system of wards that allow for a broad participation of the people.[16] As she summarizes "no one could be called either happy or free without participating, and having a share, in public power".[17]

The differences in the treatment of participation become especially clear in Pettit's discussion of the separation of powers. Although he has several passages on the theme (e.g. R, 178), he mainly attributes the decision making to the government. The parliament is left only with the task of controlling the government without participating in governmental affairs (R, 185). This treatment of the separation of powers suits Pettit because he shares the notion that a strict form of the separation of powers (which he revealingly calls 'populist') is not appropriate for a republican political philosophy. Pettit accepts that members of judiciary and executive become legislators notwithstanding their lacking legitimation through elections compared to the parliament (R, 179/180).

Pettit implicitly accepts a gap between political officeholders and the ordinary citizen. Even if one accepts such a gap (perhaps as a tribute to *Realpolitik*) – Pettit's discussion of it is unsatisfying: He postulates a deliberative democracy (R, 188–189) without mentioning the complex ideas on the public sphere from the theorists of the deliberative democracy. The systematic element that links the forum for contestations (R, 195–196) with political officeholders is missing. Whereas there are criteria for legitimate contestations there are not any institutions guaranteeing that the contestations are not only heard and recognized, but are also part of the political decision making.[18]

[15] Furthermore Pettit claims that majorities alone do not suffice for changing a law (R, 180–182). Even if the fear of policy that harms minorities is certainly reasonable, the contrary concern is not met. This may be a great deficit because one can be certain that it is dangerous for a democracy if the majority has the feeling of not being able to influence democratic decisions.

In reference to direct democracy we have to ask, if one should not separate between how a democratic decision is made (e.g. direct vs. parliamentary) and the frame, in which it can be made (usually the constitution).

[16] Arendt (1977, 253ff).

[17] Arendt (1977, 255).

[18] Here, Pettit could adopt the solution proposed by Habermas in line with Bernhard Peters (Habermas 1998, 429–435). He proposes 'locks' (a metaphor from water transport, 'Schleusen'),

One last time we therefore formulate another variation of our main question: How can Pettit reject the central role that is given to participation in a wide range of theories commonly known as 'republican' and claiming that his theory is the 'true' republican theory?

11.4 Conclusion

The distinction between a liberal and a republican conception of freedom is sound on an abstract level. Nonetheless we must ask how this distinction would prove itself in reality: Is the concrete republican state that different from a liberal state? In answering this question, we share the concerns of Matthew H. Kramer[19] and Iseult Honohan.[20] Pettit's untypical construction of liberalism's history seems to overstate the differences between his own position and broad traditions of liberal thinking. There may be a legitimate difference between liberalism and Pettit's republicanism after all. However, the difference is elsewhere, for example in the importance of civil society, for which Pettit argues in the eighth chapter of *Republicanism*. Pettit's accentuation of civil virtues, "the need for civility" (R 245) as he calls it, is certainly an expression of a republican conviction that cannot easily be integrated in a liberal political theory. On the background of these results it seems more plausible to us to call Pettit's theory a 'republican liberalism' – a proposition from Richard Dagger,[21] with which Pettit sympathizes (OPT, 10, n. 8).

Secondly, if Pettit calls himself a republican, he has to acknowledge the great differences in this tradition. Positions not shared by him should not be simply called 'populist' and put outside the tradition without a broad treatment of the multiplicity of the tradition. This necessary discussion is missing in Pettit's reconstruction of republican thought. Perhaps it would be better to acknowledge that republicanism is a 'broad church', too?

References

Arendt, Hannah. 1951. *The origins of totalitarism*. New York: Harcourt, Brace and Co.
Arendt, Hannah. 1977. *On revolution*. New York: Viking.
Ashcraft, Richard. 1987. *Locke's two treatises of government*. London: Allen & Unwin.

with which it is secured that political institutions are responsible to the informal public sphere. One example of such a lock would be a hearing of experts.

[19] Kramer (2008, 56): "Although civic republicanism as a general political doctrine can perhaps lay claim to distinctiveness, it does not provide an analysis of the concept of freedom that goes beyond the negative-liberty approach in any significant way."

[20] Honohan (2002, 185): "Yet this account [Pettit's] is still closer to the negative than the positive conception of freedom."

[21] Dagger (1997, esp. 11–27).

Barber, Benjamin R. 2009. *Strong democracy. Participatory politics for a new age*. Twentieth-anniversary edition, Berkeley: University of California Press.

Dagger, Richard. 1997. *Civic virtues. Rights, citizenship and republican liberalism*. New York/Oxford: Oxford University Press.

Filmer, Sir Robert. 1949. The anarchy of a limited or mixed monarchy. In *Sir Robert Filmer: Patriarcha and other political works*, ed. Peter Laslett, 275–313. Oxford: Oxford University Press.

Habermas, Jürgen. 1998. *Faktizität und Geltung*. Frankfurt/Main: Suhrkamp.

Honohan, Iseult. 2002. *Civic republicanism*. London/New York: Routledge.

Kramer, Matthew. 2008. *The quality of freedom*. Oxford: Oxford University Press.

Laslett, Peter. 1988. Introduction. In *John Locke: Two treatises of government*, ed. P. Laslett, 3–126. Cambridge: Cambridge University Press.

Macpherson, Crawford B. 1962. *The political theory of possessive individualism. Hobbes to Locke*. Oxford: Clarendon.

Martí, José Luis, and Philip Pettit. 2010. *A political philosophy in public life. Civic republicanism in Zapatero's Spain*. Princeton: Princeton University Press.

Pettit, Philip. 1999. Republican freedom and contestatory democratization. In *Democracy's value*, ed. Ian Shapiro and Casiano Hacker-Cordon, 163–190. Cambridge: Cambridge University Press.

Chapter 12
Focusing on the Eyeball Test: A Problematic Testing Device in Philip Pettit's Theory of Justice

Frieder Bögner, Jörn Elgert, and Carolyn Iselt

12.1 Introduction

In Chap. 2 of *On the People's Terms*, Philip Pettit discusses his republican theory of social justice which targets freedom as non-domination (OPT, 94),[1] as it is developed in the first chapter. According to Pettit, this ideal of non-domination is relevant regarding two central questions in political philosophy. The first deals with the issue of political legitimacy, i.e. the relation between a state and its citizens. We do not discuss this question here. Instead, we focus on the second: "[W]hat decisions or policies should the state impose in order to establish social justice in the relationships between its citizens?" (75). In order to answer this question Pettit develops a testing device which he calls the "eyeball test". If people can look each other in the eye without reason for fear and deference, Pettit's idea of social justice is realized. The applicability of the eyeball test is the key notion we are going to question.[2]

Within the debate of social justice, Pettit asserts two points of general agreement which are commonly recognized by political philosophers (77f.): (1) Any theory of justice faces the problem of how a state ought to organize competing claims of citizens. (2) Any plausible balance between these claims calls for an equal treatment of the citizens by the state. That the state should be "expressively egalitarian in this sense" (78) is Pettit's normative assumption.

[1] All subsequent quotations refer to Pettit's "On the People's Terms" (quoted as OPT).

[2] In political philosophy it is a common strategy to conduct thought experiments in order to establish normative principles, e.g. Rawls's original position. Even though Pettit briefly discusses his overall approach of social justice in comparison to Rawls's (121f.), it is an open question whether and in what way the eyeball test is an alternative to Rawls' thought experiment. In this paper, we do not deal with this issue. (We would like to thank David Schweikard for making us aware of this point and for many other helpful comments.)

F. Bögner (✉) • J. Elgert • C. Iselt
Philosophisches Seminar, Westfälische Wilhelms-Universität, Münster, Germany
e-mail: frieder.boegner@uni-muenster.de; jelgert@gmx.de; carolyn-iselt@gmx.net

S. Derpmann, D.P. Schweikard (eds.), *Philip Pettit: Five Themes from his Work*,
Münster Lectures in Philosophy 1, DOI 10.1007/978-3-319-26103-4_12

Within the egalitarian thesis, theories take a different stance on what Pettit calls the "substantive" and the "methodological issue" (79). The substantive issue concerns the goods in relation to which citizens should be treated equally. How equal treatment can be adequately realized with regard to any good is the methodological issue. To deal with the substantive issue, Pettit suggests that republican theories endorse "some version of freedom as non-domination [as] the good with respect to which the state is required in justice to treat its citizens as equals" (81).

Since freedom as non-domination is a crucial theme in Pettit's republican theory of justice, we will sketch its definition before we can turn to our analysis. Freedom as non-domination includes "freedom over a common range of choices" (26) and, for its realization, one must prevent the actual interference in one's range of choices as well as the possibility to intervene in someone's will (60f. & 63f.). This prevention states that it is not enough to have a superior person not actually exercising his power over his inferior.

Freedom as non-domination has to take into account two requirements: First, Pettit defines domination as "arbitrary interference" or "uncontrolled interference" (58). For this reason, during controlled interference, the interferer does not act like a "master" more than a "servant" (57). Controlled interference is given in cases when the superior agent prevents the inferior from acting towards short-term goals in accordance with the latter's "longer-term will" (57), which has priority. The interferer does not impose his own will, since he acts on behalf of the subject, who is, therefore, in control.

The second requirement concerns "domination without interference" (58f. & 63). As already mentioned, even the option to interfere influences the decisions of someone of inferior status: "The accessibility of the interference will signal your dependence on my will quite independently of the probability of my actually interfering" (60). Pettit claims that there could be no "benign" (59) mastery: An inferior person cannot look the other person in the eye without the fear that his superior might change his mind although he has never exercised his power at least not in a negative way. The subject always considers the master's opinion and reaction – she does not want to act or decide against their master's will. Even if the master does not interfere, the dominated person would not have "enough breathing space to [...] act as [he] will" (66).

Let us return to the normative assumption: Pettit claims this egalitarian thesis to be taken into account by his republican theory of social justice, which requires people to enjoy freedom as non-domination in relationship to one another (81). This notion of freedom is inspired by the image of the "liber" (82) – free person – developed by the republican tradition. It serves as a heuristic to approximate an ideal. Pettit assumes the image comprising of three lessons: The first aims to determine *what* ought to be safeguarded in a society, the second explains *how* the safeguarding is possible (83f.), and the third is the eyeball test – a controlling device – which "identifies the criterion for determining what is enough by way of safeguarding" (84). Pettit suggests that what should be entrenched are the basic liberties defined by a range of choices that are to ensure an equally non-dominated status for each citizen of a given (republican) society (83).

We assume that the first lesson is designed to provide an answer to the substantive issue, while the methodological question concerning adequate means is answered by the second lesson. This lesson about the way of entrenchment calls for juridical laws and moral values as safeguards for the basic liberties. These need to be entrenched to the point where acting against the norms is made so undesirable that there is no reason – neither subjective nor objective – for fearing interference in one's freedom of choices (72 & 83). The third lesson is an indicator of whether the first two criteria are met in a specific case: whether a constitution passes what Pettit calls the eyeball test (84).

Passing the eyeball test means that citizens "can look others in the eye without reason for fear or deference that a power of interference might inspire; they can walk tall and assume the public status, objective and subjective of being equal in this regard with the best" (84). Furthermore, passing the test includes the inability of others to interfere arbitrarily in my own affairs (84f.). Thus, the eyeball test is passed if certain basic liberties are entrenched in such a way that citizens enjoy a non-dominated status. The interdependence of the three lessons is clear: If choices happen to be entrenched as basic liberties that entail domination, and/or if the laws and norms fail entrenching the desired basic liberties, the eyeball test not only indicates domination in given cases, but furthermore "plays a part in determining" (86) what choices should be safeguarded and how they ought to be secured.

12.2 Five Possible Applications of the Eyeball Test

We now turn to our criticism, which concerns the possible applications that Pettit ascribes to the eyeball test. We identify five ways of its application and then approach them in three lines of critique.

(1) It can be used for evaluating whether the safeguarding of basic liberties, i.e. norms and laws entrenched to ensure an "un-dominated" (82) status of a state's citizens, are sufficient for that purpose (84).
(2) The test determines the potential dangers of domination in a state and how to act against these in each individual case to ensure protection of the choices that are necessary for freedom as non-domination (72 & 84f.).
(3) Furthermore, the eyeball test can be used for identifying an ideally just society and
(4) for assessing how actual states fare in regards to their approximation to this ideal,
(5) which also enables comparison of different states with each other (124f.).

In the following, we will interpret what passing the eyeball test means and will analyze its postulated applications in the light of our interpretation. Our first critique concerns the first usage. The second application will be discussed in the second and third critique, whereas we will treat the applications three to five in the third critique only.

One general assumption in our analysis is that the eyeball test is a metaphor. If it were not, in order to pass the eyeball test, people would need to have the ability to actually look others in the eye, walk tall, etc. Disregarding pragmatic problems of the measuring process due to impracticability (e.g. gathering information about each citizen's ability to walk tall, etc.), it would be hard to find a basis for evaluation (e.g. how to measure the ability to walk tall). This understanding of the test would be downright absurd. Therefore, we understand what Pettit calls "satisfy[ing]" (84) the eyeball test as a metaphoric description of a non-dominated status (95). We offer a non-metaphoric reading: If the basic liberties agreed upon in a state and the norms and laws which are supposed to be their safeguards are entrenched – in such a way that arbitrary interference in another citizen's choices is impossible –, a constitution passes the eyeball test. Consequently, the eyeball test is a mere paraphrase of Pettit's normative egalitarian assumption. Understanding the eyeball test in this way, one needs to determine whether the mentioned possible applications are actually possible.

12.3 First Critique: Evaluation of the Basic Liberties

Our first critique concerns the application of the eyeball test as the evaluation of the basic liberties and their entrenchment in order to ensure a non-dominated status. Pettit's metaphorical description of the test contains the implicit assumption of an intuitive understanding of what looking someone in the eye without fear entails. Given our interpretation of the metaphor, the test indicates that the entrenchment of the basic liberties via laws and norms is sufficient for non-domination if arbitrary interference in another citizen's choice is impossible. But the possibility of arbitrary interference is Pettit's definition of domination. Hence, the criterion for what is enough for safeguarding freedom as non-domination is not to allow domination. The result of the test is what it requires: the impossibility of arbitrary interference. Consequently, we think that the first application cannot fulfill its supposed role.

12.4 Second Critique: Potential Dangers of Domination

Our second critique affects the second application of the test mentioned above. By ensuring a non-dominated status of the citizens, the eyeball test must effectively guard against potential threats to the un-dominated relation between the people. In any practical application this task is certainly not separable from the evaluation of the basic liberties and their entrenchment to ensure the desired status (i.e. from the first application). However, the aspect of identifying dangers and working against those threats can be analyzed as a distinct element of the eyeball test: If one goal of the evaluation of the laws of a given state is to "effectively counter domination" (84), one needs to know the means necessary to achieve this end. According to

Pettit, the test informs about the required steps: "It will provide the required benchmark for adequacy in protection" (72). In our interpretation this constitutes the second application of the test.

Similar to our first criticism, it seems difficult to determine the underlying contents of the metaphor, when Pettit claims that "[t]he test suggests that you should have access to such a level of resources and protections […] that enables you, by local standards, to look others in the eye without a reason for fear or deference" (72). The alleged criteria to specify the level of protection is the notion of non-domination. Thus, we think the second application of the test pursuits a clear goal. It does not, however, inform about the means required to realize this goal.

Admittedly, Pettit mentions two criteria that we assume to serve as possible orientation in identifying the ways, in which non-domination might be curtailed. He identifies two possible variations of the capacity to interfere arbitrarily when discussing the second application of the eyeball test: "[…] the test requires that the level of protection provided in any area of choice should increase with the seriousness of the interference against which protection is needed and with the ease of access to that level of interference" (85). We interpret accessibility and seriousness of interference as useful criteria when conducting the second application of the test, i.e. when determining what potential threats to non-domination are and how to counter these, but they do not constitute the complete answer the test is supposed to provide. Towards the end of our third critique we will come back to these criteria when identifying cultural variation as a background to Pettit's concept of domination.

12.5 Third Critique: Tension between Ideal and Cultural Variation

As stated earlier, in the third critique we focus on application three to five: We want to show that there are two problems in Pettit's description of the application of the eyeball test. The first deals with a tension emerging in his elaboration of two aspects of its application: In one case, societies can be compared with an ideal, in the other case the measurement of social justice underlies cultural variation (84f. & 124). The second problem deals with the question whether the definition of "freedom as non-domination" becomes disturbed by adding the notion "rational fear" (85), which is based on cultural variation, to his concept of domination.

12.5.1 The Notion of Fear as a Problem for the Measurement of Social Justice

To discuss the first problem, we want to take a further look on Pettit's claim that with the help of the eyeball test, one can theoretically identify a "perfectly just society" (124) and that this ideal version can be used as a standard for real societies (124). In addition, one can establish a ranking which displays the order of the societies' approximation to the ideal version (124).

In contrast to what Pettit says in this passage (124f.), he asserts in a second one (85) that there is no "universally valid alternative" to "local standards" which determine benchmarks for "when fear and deference is irrational and when prudent" in *one* society (85). Thus, local standards, instead of an ideal, determine the benchmark for evaluating social justice. The eyeball test is passed if people "can look others in the eye without fear or deference" (84). We interpret "rational fear" as that kind of fear which stems from the justified belief that other people are able to interfere on an arbitrary basis with my options. In contrast, when applying the eyeball test "mere timidity or cowardice" (84) shall not be taken into account. Pettit claims that there are "local standards" which fix the parameters that are used to weigh rational fear and therefore to reveal domination – i.e. the benchmark shows the limit of rational fear and indicates when the eyeball test is not passed. He explains that fear or cowardice depend on "cultural variation in what counts as mere timidity rather than rational fear or deference" (85).

But if one is always obliged to find out the local standard for rational fear in order to identify a specific benchmark of domination for each society, can one still compare different societies? We think that the assertion, that one needs an ideal of a just society to evaluate real societies, and the requirement of local standards get in conflict. After introducing an ideal by which actual societies should be judged, Pettit admits that for the application of this contrasting juxtaposition, we need more detailed "stipulations" (125) for taking into consideration the individual characteristics of each society. Different societies may realize social justice in distinct ways, but "perhaps quite intuitively", they implement the same "degree" (ibid.) of social justice. This variation underlines the difficulty in establishing one ideal with universal validity.

12.5.2 The Notion of Fear as a Problem for the Definition of Freedom as non-Domination

In the following, we deal with the second problem of our third critique. We want to show that the tension between the ideally just society and the cultural variation of fear can be supported by another argument: Pettit defines the terms 'freedom as non-domination' and 'domination' in the first chapter. We want to reveal that these definitions become relative when, in the second chapter, 'non-domination' is described as a state, in which no one has to fear arbitrary interference by anybody. The fact that the understanding of fear depends on local standards disturbs the definition of domination and, consequently, the definition of freedom as non-domination.

As we have outlined in the introduction, in Pettit's notion of freedom, domination is defined by 'uncontrolled interference' or having the option to interfere anytime without having done it previously or currently doing it. When Pettit proceeds to elaborating on how his concept of freedom can be realized, his approach changes from an abstract account to a more practical domain:

> With the argument for this understanding of freedom in place, we can go on in the next chapter to ask after what it would mean to establish equality for people in the enjoyment of freedom: that is, to ensure an equal status for them as free citizens. This ideal amounts, as we shall see, to a republican ideal of social justice (28).

In the first chapter, Pettit equates the range of choices, to which everyone must have free access, with the basic liberties and he ascribes the evaluation, whether the basic liberties as well as their entrenchment by laws and norms are enough for freedom as non-domination, to the eyeball test (47). When he first mentions the eyeball test in *On the People's Terms*, he relates it to local standards which are used as "yardstick[s]" (47) to determine when people can "look one another in the eye without fear or deference" (ibid). Pettit assumes that the subjectivity of the test is not of great concern and without further precision, he supposes an "accepted sense of where the benchmark lies" (72). Pettit looks for a yardstick which is "culturally established" (47). It is used to measure "the extent to which people" have access to the basic liberties and how "undominated" (82) the people are in the exercise of the basic liberties. After developing his abstract theory of freedom, Pettit proceeds to establishing his version of social justice which targets freedom as non-domination. An un-dominated or free status is achieved when citizens do not have warranted reasons to fear arbitrary interference.

At this point of the argumentation the passage about local standards becomes relevant (84f.). When Pettit speaks of the eyeball test, he ends his abstract theoretical treatment of freedom. Here, the test is used to evaluate concrete situations: The basic liberties and their entrenchment are realized in a satisfying way, if people are not warranted to feel the anxiety of being dominated. But in his precise description of the functionality of the test (84f.), Pettit postulates that the supposed objective benchmark underlies cultural variation. The contexts, that give rise to the feeling of anxiety, depend on cultural background and whether fear is rational is based on an "accepted sense of where the benchmark lies" (72) in this specific society. Above we have proposed that this variation of the benchmark might be opposed to a universal ideal, as Pettit points out that "there is not going to be any universally valid alternative that might be invoked in their [the benchmark's] stead" (85). The benchmark for rational fear is not universally valid.

Moreover, Pettit demands that the eyeball test "should be understood and applied" to identify variations in two kinds of measuring interference: the "seriousness of interference" and the "access to that level of interference" (85). But we assume that in different cultures, the same actions might be evaluated as "more or less serious form[s]" (85) of interference. In some societies, the fact that a father can determine his daughter's choice of a husband might be judged a less serious form of interference than in another culture. Thus, we question whether a test which is based on cultural variation can serve to identify the variation in seriousness and access.

We admit that within the conceptual analysis of freedom as non-domination a definition of fear is not necessary. The term 'fear' is not mentioned in the account of domination in Chap. 1. But when Pettit talks about the realization of social justice in the basic liberties and their entrenchment, a definition of fear is required. The eyeball test proofs the adequacy of the basic liberties and their safeguards. When explaining the test, Pettit does not just take theoretical considerations into account, because he refers to cultural variations to qualify his notion of fear. His understanding of domination, which is relevant in chapter one, is a theoretical concept. In the second chapter, the term 'domination' is closely connected with the eyeball test (the positive result of the eyeball test can be interpreted as a metaphor for an un-dominated status). But in order to examine whether freedom as non-domination is realized, one cannot treat the test as a mere theoretical construct. To reveal whether people can look one another in the eye without fear, one needs to know what fear is and whether it is rational. Because of the close connection between domination and the eyeball test in chapter two the clear definition of domination becomes disturbed by the term 'fear', which is substantial for the eyeball test, apparently culturally variable and lacks a definition by Pettit. Consequently, this problem also concerns the definition of freedom as non-domination, because it might become a relative term: It is always relative to a certain understanding of fear which indicates the existence of domination in a society with a specific cultural background.

Therefore, one can question whether Pettit's demand at the end of the second chapter is reachable: Can the eyeball test be used to establish an ideal version of social justice which is simultaneously applicable to evaluate real societies (124)? One can accept the image of the eyeball test as an explanation of an un-dominated situation and perhaps it can serve purposes similar to a thought experiment: After identifying basic liberties and constituting laws, one can theoretically go through the situation of the eyeball test. But by making it relative to different cultures, Pettit disturbs the definitions of his basic concepts and produces the tension between cultural variation and his suggestion to use the eyeball test to identify an ideal which could be the "goal" (123) that every state should promote.

12.6 Conclusion

In conclusion, we want to highlight the results of our analysis. We have identified five possible applications of the eyeball test. For each usage, we have discussed problematic aspects. In the first critique, we came to the conclusions that Pettit's criterion for what is enough to ensure freedom as non-domination is not to allow domination and that the first application can thus not fulfill its assumed role. In the second critique, our result was that the second application of the test on the one hand has a clear goal – identifying threats of domination and possible counteractions –, on the other hand, it does not inform about the means required to realize this goal. The third critique concerns applications three to five. We discussed two problems in Pettit's description of the application of the test: First, there is a tension between the ideal in light of which societies can be compared with one another and the cultural variation underlying the measurement of social justice. Second, Pettit's idea of freedom as non-domination becomes disturbed by adding to his concept of domination a notion of fear, which is based on cultural variation.

We would like to make a suggestion that exceeds some of the findings of this paper. One can ask whether the problems of the eyeball test presented might constitute more substantial issues which go beyond the test as an inspection or examination procedure but concern the basic concepts of Pettit's theory of political freedom: If the eyeball test lacks proper guidelines to fulfill its supposed role – as shown by critique one and two – the concept of republican freedom in this theory has no suitable testing device for the basic liberties the theory aims at. Furthermore, if the notions of justice as an ideal with universal validity and freedom as non-domination have culturally relative underpinnings – as shown by our third critique – the republican idea of freedom and justice presented in *On the People's Terms* needs further refinement.

Part III
Reply Essay

Chapter 13
Self-defense on Five Fronts: A Reply to My Commentators

Philip Pettit

It is very challenging to be presented with such a wide array of commentaries on my work, many of which raise very fundamental questions. In this necessarily brief set of responses, I propose to divide the commentaries into five areas; to present my overall background view in each area; and then, against that background, to consider the main points made in the relevant commentaries. The five areas are: Epistemology and Semantics; Philosophy of Mind; Consequentialism; Group Agency; and Republicanism.

13.1 Epistemology and Semantics

The two papers of relevance in this area are 'Rule-following and A Priori Bi-conditionals—A Sea of Tears' and 'Pettit's Mixed Causal Descriptivism: Feeling Blue'. Together the papers raise issues about my account of how the basic terms or concepts with which we operate—those that are not defined for us in other terms—get to have semantic values; about the knowledge which that mode of semantic connection is meant to give us, as an a priori matter; and about the claim that while having a descriptive character, such knowledge does not amount just to knowledge by description as distinct from knowledge by acquaintance.

Both papers go back to my more or less Wittgensteinian views as to how we succeed in being able to gain access to the property that a predicate is used to ascribe—say, the property of regularity that we use the term 'regular' to ascribe to certain figures—and to gain access, more generally, to the meaning of any basic term in our

P. Pettit (✉)
University Center for Human Values, Princeton University, Princeton, NJ, USA

School of Philosophy, Australian National University, Canberra, Australia
e-mail: ppettit@princeton.edu

S. Derpmann, D.P. Schweikard (eds.), *Philip Pettit: Five Themes from his Work*,
Münster Lectures in Philosophy 1, DOI 10.1007/978-3-319-26103-4_13

language.[1] By hypothesis, we do not learn the meaning of ‘regular’ by definition in other terms; such a definition may be available in geometry but as ordinary users we master the term in a more direct, ostensive manner. So the question is: how can we know what property it is that the predicate is used to ascribe?

The question is daunting because of the observation that Wittgenstein puts in place, which is that no finite set of examples could provide the means of cottoning on to the property that we intend the predicate to ascribe. Consider examples such as these:

No such finite set of examples could direct us to the class of instances, indefinite in extension, that corresponds to the property of regularity; after all, there is nothing to determine that the right way to go on from the finite set would take us to all the other instances. And so it seems that we cannot learn the meaning of ‘regular’—that is, cannot grasp the property the predicate is supposed to ascribe—on the basis of ostension. Indeed, to take Kripke’s way of highlighting the problem, the fact that in the past I used that set of examples to try to determine the meaning of the term would not be enough to assure me now—assuming, that I do not now face a similar difficulty—that it was indeed regularity that I meant then to pick out.

In response to this problem I developed a story that involves a number of distinct stages.

- In virtue of our biology and training we naturally develop an extrapolative disposition when presented with a set of examples like the examples of regular shapes; we form a more or less brute disposition to go on in a certain way, casting some shapes as relevantly similar and others as dissimilar.
- We do not mindlessly give expression to this disposition, however, as Kripke suggests we might do. Rather we use the disposition to present to us, at a first pass, the indefinite class of instances—and so the property—that we have in mind: it is *that* property, we think, the one exemplified by the examples for those of us who share the disposition.
- Assuming that the disposition is indeed shared with others we baulk at any discrepancy of response, as when you extrapolate in one way, I in another. And we thereby display a presumption that we are each responding to the pressure of a common constraint: the cues provided by a property that we are tracking in common.
- We resolve discrepancies, at least in the ideal case, by identifying factors that explain it. Thus we might explain the fact that you treat a certain new shape as similar to the initial examples, I as dissimilar—you treat it as regular, I as

[1] For an account of the views relevant to this section see CM as well as (Pettit 1990a, b, 1991c, 1998a, c, 1999b; 2002b; Jackson and Pettit 2002).

irregular—by the fact that, as it turns out, I can be shown to be wearing distorting spectacles.

- Among the potential explainers of any discrepancy we treat those explainers as best that fit most naturally with the presumption that there is an objective property that we are each tracking; we take ourselves to be triangulating on that property, not just coordinating for coordination's sake. Thus we might well decide in any discrepancy that the majority are wrong, on the grounds that it is they who are wearing distorting glasses.
- The upshot is that we can each identify the semantic value of a basic predicate like 'regular' insofar as two things hold: first, we take it to be presented to us via a disposition that is subject to correction in constrained adjustment to the dispositions of others; and, second, people's dispositions prove relatively convergent when they operate under the constraint of triangulating adjustment.

This story of how we gain semantic competence in the use of a basic term sounds very intellectual but it is not necessarily so. It does not say that this is how we think—that this is how the child thinks, for example, in learning the meaning of 'regular'—but only that it is as if we think like this. Were we able to think like this then we would be able to be as determinate in the properties we mean to track with our basic predicates as our community is determinate over time in the convergence of associated dispositions. The fact that it would be enough to be able to think like this means that there is no longer a mystery associated with rule-following: that is, with following or tracking a property like regularity that is indefinite in extension. We bring off that achievement in virtue, as Wittgenstein would say, of relying on a custom or practice in which we participate with others.

On the basis of this story about rule-following I argued that our semantically basic concepts are all, in a sense, response-dependent. In order to master or possess those concepts, we need to be creatures to whom certain responses—in our example, extrapolative and adjustive dispositions—come naturally. The concept of 'regular' is not response-dependent in the sense of being defined as that property, whatever it is, that gives rise to the corresponding extrapolative disposition; in this way, it contrasts with a concept like 'disgusting'. But it is still a concept that no one who lacked our natural extrapolative and indeed adjustive dispositions would be able to grasp and master in a non-parasitic way.

The fact that it is a response-dependent concept of the kind explained means, I argued, that a certain a priori bi-conditional holds true, as is generally said to be the case with response-dependent concepts. It is a priori knowable for us as philosophers or commentators that something is going to count as regular for participants in the practice if and only if it looks regular to them—it looks to be of a kind with the initial exemplars—in the absence discrepancy-explaining factors: that is, as we can say, in normal conditions. This is a priori knowable—that is, knowable without recourse to empirical testing or investigation—insofar as the argument just given, which is conducted on an a priori basis, is actually sound. But more important that this, it is also going to be a priori knowable for the participants in the practice that something is regular if and only if, in my sense, it looks regular under normal conditions; such conditions will be salient for them as conditions in which discrepancy-explaining perturbers are absent. This bi-conditional is going to be a priori knowable for the participants, insofar as they are capable of the sort of philosophical reflection pursued here. But the fact that it is a priori knowable, of course, does not suggest that they will actually know it: they may not conduct the reflection required.

With these comments in place, I can now address the questions raised by the authors of 'Rule-following and A Priori Bi-conditionals—A Sea of Tears'; indeed the comments have been written with a view to answering those questions. Are the a priori truths of the kind I envisage analytical truths? Not in any obvious sense. While they can be known independently of empirical testing or inquiry, they cannot be known just on the basis of an analysis or definition of the terms. That might be the case with the a priori bi-conditional that would be appropriate for a response-dependently defined term like 'disgusting': it is a priori that something is disgusting if and only if it gives rise to disgust under normal conditions (though not perhaps normal conditions, as I define the phrase). But it is not going to be true of a priori biconditionals that reflect response-dependent mastery as distinct from response-dependent definition.

Do the a priori bi-conditionals for terms like 'regular'—or any other basic terms like, for example, 'red' or 'blue'—support a sort of infallibility on the part of users? Yes and no. Yes, in the sense that if something looks regular or red under normal conditions then that is an infallible guarantee that it is regular or red. No, in the sense that there is never an infallible guarantee that conditions are in my sense normal; there is always the possibility that adjustment across time, whether on the part of an individual or a whole community, may lead to characterizing conditions of an earlier judgment as less than normal: as conditions in which a certain perturbing factor was at work.

Finally, what exactly do my a priori bi-conditionals relate? In the works on which the authors of the paper were commenting I did not make clear that some truths are a priori knowable for commentators and other related truths for participants and I am grateful for having been pressed to be explicit about this. I am also grateful for the work of the authors in revealing an obscurity in some formulations of my claim about the truth that is a priori knowable for participants. I suggested in some places that what is a priori for them is not that something is red, but only that it is denominably red, if and only it looks red in normal conditions. In making this move, I wanted to emphasize the basis in response-dependent mastery rather than response-dependent definition of the a priori truth in question. But this led to confusion on my part. An implicature or presupposition of the participants' saying, in their language, that something is red under such and such a condition is that they can name the property in their language. The supposedly cautious formulation in terms of what is denominably red is unnecessary and misleading.

What now of the challenges raised by the authors of the second paper 'Pettit's Mixed Causal Descriptivism: Feeling Blue'? They address my attempt to tell a story for how a basic predicate like 'blue' might come to have its meaning, according to which we, the users of the term, will identify the property at issue in a quasi-descriptive manner. In our minds, so the story suggests, it will be that actual property, the one that is ostensively available in such and such examples and makes things look blue for us, at least when no independently identifiable perturber is getting in the way. The property will be associated in our minds, then, with a general descriptive characterization, as descriptivism assumes, but we will endorse that characterization on the basis of recognizing examples of blueness and recognizing that it applies to them. We will know that blueness is the ostensive property that makes things look blue in the absence of pertubers on the basis—and perhaps only on the basis—that for everything that we take to be blue we expect it to display that character.

I told this story in order to defend a sort of descriptivism—specifically, a rigidified descriptivism in which 'blue' picks out the actual property that makes things look blue—against some challenges raised by Robert Stalnaker. I maintained that if the basic terms are associated with such anchored descriptions, as I called them, then there is not going to be a permutation problem of the kind that is often thought to undermine a global descriptivism. More importantly, because more controversially, I also held, first, that on the story told I can avoid the holistic slippage that would make it unclear which descriptions are privileged in determining the semantic value of 'blue'; and second, that equally I can avoid the indirectness charge that 'blue' refers in the mouths of users to whatever property, they may know not what, that meets the condition. The slippage is avoidable insofar as the description on the right-hand side of the relevant bi-conditional is naturally privileged. And the indirectness is avoidable insofar as the description is endorsed in a manner that presupposes acquaintance with specific instances of the blueness property.

The authors of 'Pettit's Mixed Causal Descriptivism: Feeling Blue' argue against the general line just described on the basis of a principle I reject. They assume that if it is a priori knowable that p if and only if q, then it must be the case that the expressions on either side of the bi-conditional are synonymous: that 'they have the same meaning, i.e. the same descriptive or informative content'. And then they use that assumption to reject the a priori status that I give to the statement: 'Blueness is that ostensive property that makes things look blue in normal conditions'. Their assumption would be reasonable with a response-dependently defined term. Thus the a priori status of 'X is disgusting if and only if it occasions disgust under normal conditions' would support the corresponding synonymy claim. But the assumption is certainly not going to hold with the a priori bi-conditional that goes with response-dependently mastered terms such as 'blue' or 'red' or 'regular'. And so I do not think that it can be used to undermine the position that I defend against Stalnaker.

13.2 Philosophy of Mind

Issues in the philosophy of mind come up in three papers, 'Playing Pong with the Mind? Pettit's Program Model and Mental Causation', 'Discovering the Properties of "Qualia" in Pettit's Theory of Phenomenal Consciousness' and 'Notes on Pettit's Concept of Orthonomy'. Again, let me set out some relevant background and then address the challenges raised in these papers.

The main element in the background is the physicalist and broadly functionalist theory of mind that I adopt. According to the physicalism I embrace, the actual world is such that if we could replicate it in all physical respects—say, in all microphysical respects—then without doing anything else in addition, we would still manage to replicate it in all respects whatsoever, including all mental respects. We or exact counterparts would exist in the physical replica of the actual world and would enjoy the same intentional and phenomenal mentality that we enjoy in the actual world.

This is to say that the way the actual world is in all respects, physical and non-physical, is supervenient on the way it is in physical respects. The physical characteristics of the world determine its non-physical characteristics in the way in which the positions of the dots in a pointilliste representation determine the figures and shapes in the painting. There may be variations in the physical characteristics that preserve the non-physical features of the world, as there are variations in the dots that preserve the figures and shapes of the painting. But there is no way of varying those non-physical features while the physical characteristics remain the same, as there is no way of varying the figures and shapes while the positions of the dots remain the same.

A good way of explaining the supervenience entailed by physicalism is to adopt a broadly functionalist account of most of the non-physical features of the world. On that account a feature such as my believing that something is the case—say, that p—consists in my instantiating a physical state—say, a neuronal state—that plays a certain role or function. When independently defined normal conditions obtain, that state appears or survives in the presence of evidence that p, disappears in the presence of evidence that not-p; it combines with any state that counts by its role as the belief that if p, then q to generate inferentially the belief that q—or perhaps to undermine either the belief that p or the belief that if p, then q; and if it identifies a way of X-ing, then it combines with the desire to X and the belief that there is no other way to X to generate an action of X-ing. If an intentional property such as believing that something is the case is a functional property of this kind, then it is easy to see that that property can be realized by the presence of a physical property that plays the appropriate role. This is a condition that can be realized for one intentional property only insofar as it is realized simultaneously for many—the point should be obvious from the observation about inferential role—but that is no objection to the account.

If all the non-physical properties in the world are functional properties, or are properties that can be realized in the same straightforward way by physical states, then it follows that the world displays an architecture in which properties—and any entities like events or processes or substances that are individuated by properties—appear at different levels of composition. Any higher-level properties and related entities will be immediately supervenient on properties at a lower level and those lower-level properties and related entities may be supervenient on properties at a lower level still. And so on to a presumptively last level of microphysical properties.

This picture raises three questions, which are addressed respectively in the three papers in philosophy of mind. The first question is whether this hierarchical picture allows us to think that higher-level properties such as the intentional property of believing or desiring something can enjoy causal relevance in any sense. The second is whether it allows us to make sense of the phenomenal states of consciousness that we enjoy when we see and savor a look or taste or smell: when there is something distinctive and ineffable that it is like to undergo such an experience. And the third is whether it makes room for a significant conception of freedom and autonomy. I shall discuss each of these issues in turn.

I assume that whatever causal relevance a property—or the instance of a property—enjoys at a given level Ln is dependent on the causal relevance of the properties—or property-instances—at the immediately lower level, Ln-1. But that suggests that all the causal power of the higher-level property belongs in truth to the lower-level properties, whose power may in turn belong to properties at lower levels still. And so it may seem, implausibly, that there cannot really be higher-level causation at all. Thus, to take a strikingly implausible example, there cannot be any causation involved in the relation between my beliefs and desires and the actions I perform.

In order to account for why we do readily ascribe causal relevance to higher-level properties like beliefs and desires, Frank Jackson and I introduced the program model of causal relevance.[2] This model was meant to be an improvement on a model of supervenient causation introduced earlier by Jaegwon Kim (1984). In Kim's view we can say that A is a supervenient cause of B just in case *alpha,* the configuration of properties or property-instances that realizes A, is a cause of *beta*, the configuration that realizes B. That view suggests that higher-level causes are insignificant and that learning that A causes B gives us no information about the causal linkage over and beyond the information we would have if we knew that the *alpha* configuration causes the *beta* configuration.

According to the program model a higher-level factor like A can count as the cause of a higher-level factor B only if a further condition is fulfilled. Not only must the actual realizer of A be the cause of the actual realizer of B. It must also be the case that no matter how A was realized at the lower level—at least within certain bounds—the realizing configuration would give rise to a configuration that realized B. In such a case we might say, in the metaphor we used to describe the model, that A programs for B, where the programming is implemented by the lower-level configurations. A programs for B in the sense that its presence means that things are so

[2] For our joint papers on the topic see (Jackson et al. 2004). And for work of my own on the topic see (CM; JD).

organized at the lower level that they are more or less bound to give rise to the presence of B.

David Lewis (1986) characterized causal explanation as an account that gives information on causal history and good causal explanation as an account that by contextual criteria gives good or useful information about causal history. There is no doubt but that learning about the programmer of an event involves gaining information about the causal history of the event that is of the first importance in many enterprises. Knowing that A programs for B gives us information that we might not have had, if we knew just that *alpha* caused B. To know that A programs for B is to know that however A actually gave rise to B—as it happens, by means of the linkage between *alpha* and *beta*—it would have given rise to it even if the linkage had not been of that kind. It is to have modal information about that history, being able to recognize, for example, that if you want to bring about B you need only worry about bringing about A; let A occur and, no matter by what lower-level route, B is bound to follow.

A is causally relevant to B in the straightforward sense, then, that having information about the A-B connection is having information about the causal transaction involved that you might not have had just by knowing about the *alpha-beta* connection. Hence there need not be any hesitation about ascribing causal relevance to programming factors as well as to the lower-level factors that implement the programming, whether in an intermediate or ultimate fashion. And this is true no matter how we choose to define a cause. The program model offers a story about how causal levels relate to one another that holds under any of a variety of accounts of what it takes to be rightly describable as a cause.

I think that the program model applies in a large range of cases. In particular it applies in the psychological case of intentional properties like beliefs and desires causing actions, where the causal linkage involved must depend, at least according to physicalism, on the lower-level role of neural configurations in causing actions. The model may not be fully satisfying for those who worry, as Jaegwon Kim worries (1998), about the significance of higher-level causation.[3] But it ought to make clear that higher-level causes do not come cheap—they require a programming architecture—and that they are worth the price we have to pay: they provide potentially vital information on causal transactions.

The authors of ‘Playing Pong with the Mind’ do not agree. They invite us to consider a simple game in which by moving a racquet—or rather the image of a racquet—on a computer screen, I can hit a ball—or rather the image of a ball. They appear to assume that in such an example it need only be the case that the electronic realization of the movement causes the electronic realization of the impact on the ball. And under that assumption they suggest, reasonably, that ‘the electronic story taking place in the processing hardware of the computer’ gives the lie to ‘the graphical story taking place on the monitor’. But it should now be clear how I am bound

[3] Notice that when Kim (1998, 74) considers whether the program model ought to relieve that worry, he mistakenly suggests that that was the model he had proposed in putting forward his views on supervenient causation in his 1984 paper.

to reply. Suppose that the computer is complex enough—as plausibly it must be—to make it the case that no matter how the movement of the racquet is realized electronically it will give rise to an impact on the ball, no matter how that is realized electronically. Suppose, in other words, that it is complex enough for it to be the case that the movement programs for the impact. In that case, I argue that it makes perfectly good sense to say that the graphical movement is a cause of the graphical impact. In other words, I bite the bullet.

The second question raised by my physicalist, broadly functionalist view of the universe, which is addressed in the second paper under consideration, is whether there is any room under that approach for phenomenal consciousness: that is, for the looks and tastes and smells, for example, that are hailed as instances of 'qualia'. This is a deep and troubling issue and what I have to say is somewhat tentative, drawing on a range of papers I have written over the years.[4]

Imagine that Eva occupies a stroboscopic room much as Mary, in Frank Jackson's thought experiment, occupies a black and white room. The stroboscopic lighting in the room means that Eva does not perceive motion in the ordinary way; she sees objects jump from place to place but does not see them move in the ordinary sense of the term. Like Mary, Eva has knowledge of all the physical facts of the universe, understanding that motion is continuous change of place and being able with the help of fancy computing techniques to determine the direction, velocity and acceleration of any moving object. Now imagine what it must be like for Eva when she leaves the stroboscopic room and suddenly has the capacity, just by looking, to see the various aspects of moving objects: to see where and how fast they are moving, for example, and to reckon in a flash where she should reach in order to grasp a moving object that is passing by.

I assume that on leaving the stroboscopic room Eva would enjoy a transformation in her experience of the sort that we associate with phenomenal consciousness. There would be something it is like to see an object moving—there would be a look that motion assumes—and this is something that she couldn't grasp from within the stroboscopic room. Yet few of us will be inclined to think that this look of motion is a special non-physical feature of the universe that erupts into existence. For all that has happened with Eva is that she has learned a new, experiential skill in detecting a property that she always understood fully and was able to detect in a less direct manner. There is an appearance or look of motion that goes with exercising that skill. But there is no reason to think of this look as anything over and beyond the representation of the motion in someone who has such a skill.

Of course Eva's representation of the motion, once she has left the stroboscopic room, is not just a representation in her brain of which she may have no awareness. It is a representation to which she can pay attention, asking about whether it is misleading or not: misleading, for example, in the way in which the appearance of moving is misleading when you are on a stationary train and another train moves beside you. And so the representation is, as we may say, a representation *for* Eva, not just a representation *in* Eva. But still the representation is a resultant of her new

[4] Relevant papers include (CM; 2003c, 2004, 2007a; JD; 2008c).

skill at identifying an old and familiar property. It appears superveniently on the presence and exercise of that skill and not in the fashion of a new existence.

Wittgenstein once asked Norman Malcolm why people would ever have thought that the sun crosses the sky rather than that the earth rotates. Malcolm replied: because it looks that way. To which Wittgenstein responded with a question: what would it look like if it looked like the earth rotates? We might grant in Eva's case that it looks like there is a new, non-physical property available in her experience outside the stroboscopic room. But we should not derive an 'is' from this 'looks'. Like Wittgenstein we should ask what it would look like to Eva if it looked like she had a new representation of the motion of objects in her visual field.

Like the authors of the paper 'Discovering the Properties of "Qualia" in Pettit's Theory', I agree that if qualia—that is, if looks, smells, tastes and so on—had an intrinsic character, then that would raise a serious problem for a functionalist physicalism of the kind that appeals to me. It would mean that there is no reason to block the thought that while everything remained the same in the functional infrastructure—in the neuronal character of the mind—still the qualia might come or go. There would be no block to the possibility of David Chalmers's zombies and no objection to thinking of the qualia as free-floating additions to the functional universe.

Drawing on the Eva scenario, however, I argue that consistently with a physicalist and functionalist picture, we can explain in principle why there might seem to be qualia of this kind. Taking the hardest case of all—the look of color rather than the look of motion—I argue that something is possible here that is exactly parallel to what plausibly obtains with the look of motion. The visual tracking of objects that is associated with color is circumstantially robust: it involves tracking them as they move or remain still under different levels of illumination, against different backgrounds, and from a position that is itself stable or moving. It seems to us that it is the look of the color—the look of that red cat, darting about the bushes—that enables us to track the object, in which case it would be hard to deny that look an intrinsic character. But why should it not be the other way around? Why should it not be the case that we enjoy access to the colored look of an object in virtue of having and exercising our circumstantially robust, visual skill in tracking it? Far from being something that guides us in tracking the object, on this account the colored look would be supervenient on the tracking skill itself. It would relate to the tracking as product or precipitate, not as producer.

The idea behind the proposal should be clear from the parallel with the motion case. Being able to detect an object's motion visually, we form a representation of it in ourselves and when we make that representation into an object of attention, we have an experience as of a motion look. Being able to track an object on the basis of vision alone, and to track it under many variations, we form a corresponding representation of it in ourselves. And when we make that representation into an object of attention, we have an experience as of a color look. The color look does not guide us in visual, circumstantially robust tracking; it is the representation that forms in us as a result of exercising such a tracking skill and that can be made itself into an object of attention.

I hope that these observations explain why I can agree that while it seems that experience provides us with intrinsically identified qualia, there is not an inescapable need to treat this appearance seriously; there is a way of explaining it away that is consistent with functionalism and physicalism. I add two more observations that I have made elsewhere. One is that there is experimental evidence that supports that explanation. And the other is that it is possible to give an account of why the belief in qualia is so hard to dislodge.

The experimental evidence derives from the work of Ivo Kohler in the 1930s. He found that wearing color-distorting glasses affected people's capacity to track and deal with their environment but that after their practical capacities began to improve—this could take up to a couple of months—their sense of color came into line at the same time. There may be a number of ways of explaining this effect but one obvious hypothesis is that color looks derive from color tracking capacities, not the other way around. And that would fit perfectly with the position adopted here.

The other observation is that we should not be surprised that it is so hard to believe that color looks are supervenient on tracking capacities. There are a number of cases where it makes good sense to believe in supervenience but where it is impossible to make the supervenience salient, as in an 'Aha' experience. You may know on an independent basis that the profile or movement of a figure on a screen is supervenient on the positions of the pixels but not be able to make this salient in experience: if you are near enough to the screen to see the pixels you will be too near to see the figure; and if you are far enough away to see the figure you will be too far to see the pixels. I suspect that something similar holds in the case of phenomenal consciousness. We can or might be able to scan the neuronal changes associated in our brain with color capacities but we could not do this simultaneously with savoring the changes in our color experience. Thus we may never be able to vindicate in experience the sort of supervenience for which I have been arguing here. It may be as difficult to see our color experience as a precipitate of practical visual ability as it is to look on the sun and have an experience as of the earth rotating.

The third paper in philosophy of mind raises questions about the theory of autonomy that I have defended, mainly in work pursued in collaboration with Michael Smith.[5] The theory by which I hold argues that we should think of autonomy as the virtue of being guided in the formation of your desires, and therefore actions, by the values you espouse and being guided in the formation of your values by whatever evidence is relevant to that matter. Smith and I described this conception of autonomy as an ideal of orthonomy: an ideal of being guided by the orthos or right, as distinct from the autos or self.

In the initial presentation of that ideal we argued that it is supported by the observation that while every intentional action is the product of beliefs and desires—you act in a way that promises, according to your beliefs, to satisfy your desires—the reason you choose to act that way is only very rarely that you desire such and such and that acting that way will satisfy your desire. I offer money to someone begging

[5] See (Pettit and Smith 1990, 1996; TF; JD).

on the street out of the belief that they need help, combined with the desire to provide help. But the reason for which I act is just that the person needs help, not that I desire to provide help and that giving money will satisfy that desire. The option that I take to be imperative is one that that reason would oblige me to take, as I actually see things, even if I did not happen to have the desire. The fact that I desire something counts in favor of an option only in the case of desires like the yen to have a smoke, or the compulsion to relieve an itch, which have a phenomenological presence; only in that sort of case does it make sense to think of my action as a way of satisfying the desire and relieving the itch.

Smith and I summed up this thought by saying that while desire is always present in the background of an agent's psychology, combining with belief to produce—strictly, program for—the action, it does not normally figure in the foreground of the agent's attention: that is, it does not often figure in the role of providing the agent with a reason to act. This strict-background view of desire, as we called it, supports many lessons that we tried to gesture at in our original 1990 paper on 'Backgrounding Desire'. And one of those lessons, so we suggested, is that it is wrong to think of autonomy as a virtue that consists in considering one's desires and acting on them only in the event of having a higher-order desire that you should be moved by them. On one interpretation, that is what Harry Frankfurt proposed in his well-known, hierarchical view of autonomy. Arguing that autonomy consists in acting on ground-level desires that you have a higher-order desire to be moved by—in later versions, a higher-order desire that you identify with—he suggested that to try to be autonomous you have to test your desires for whether they pass this constraint, thereby bringing them into the foreground.

We maintained that on the strict background view of desire, it is better to have a conception of autonomy that does not require this foregrounding of desire. We proposed that on a natural interpretation, autonomy should rather be taken to consist in forming desires that fit with your values, where the test for whether you are being autonomous consists in establishing that the options you choose do indeed display values you cherish. This is to establish, in our terms, that you are making your choices in accordance with the right; more specifically, in accordance with what you take to be right.

The authors of 'Notes on Pettit's Concept of Orthonomy' suggest that the conception of autonomy as orthonomy is not strictly entailed by the strict background view of desire and I think that is right; we suggested only that it is given a certain plausibility by that view of desire. They also raise questions to do with what values are on this approach, with whether the values relevant are meant to be subjective or not, and with how orthonomy relates to conversability. I turn now to those issues.

We conceived of values in our discussion as properties—perhaps agent-relative, perhaps agent-neutral—that, roughly speaking, you as an agent are disposed reliably to cherish; as between options that otherwise leave you indifferent, you are disposed to pick an option that has the property over options that lack it. The idea is that if you are as other human agents, you will form desires over options on the basis of the cherished properties—strictly, the reliably cherished properties—that you take those options to display; you will be sensitive to what Elizabeth

Anscombe described as desirability characteristics. The cherished properties will be your values, the despised your disvalues.[6]

Are there any constraints on the properties that it is appropriate in this sense to cherish or, as we also say, desire? In the 1990 paper we did not address this question but in the 1996 follow-up paper on ‘Freedom in Belief and Desire’ we suggested that if there are, then satisfaction of those constraints should also be a requirement of orthonomy. To be orthonomous, on this conception, is to form your desires and projects in a way that is faithful to your values and to form your values in a way that is faithful to whatever constraints—say, of consistency, evidential support or normative support—that are taken to be relevant. To be orthonomous is, on the one side, to avoid epistemic irrationality in the formation of your values and, on the other, to avoid executive irrationality—say, weakness of will or caprice—in acting on those values.

How finally does orthonomy related to conversability, a concept that Smith and I introduced in the 1996 paper? To be conversable is to be capable of being reached in the reason-giving, reason-checking exchange that often characterizes conversation. It is to have the capacity to register and respond to the reasons I can give you for desiring or believing this or that, invoking values that you have or ought under relevant constraints to have. Conversability in that sense is quite consistent with not being orthonomous, since you can have the remote capacity to register and respond to values without exercising that capacity in a reliable way. Orthonomy, by contrast, is the achievement that consists in the exercise of that remote capacity or, perhaps better, in the development of a more proximate capacity for such an exercise. On this view conversability is like the capacity that many human beings, trained and untrained, have to play the piano, whereas orthonomy is like the capacity that only those who have learned to play the piano can be said to possess.

13.3 Consequentialism

There are two papers that address my consequentialist commitments in moral and political philosophy. ‘Two Problems of Value-Monistic Consequentialism in Philip Pettit’s Theory of Criminal Justice’ looks at consequentialism in my political theory and ‘Indirect Consequentialism and Moral Psychology’ examines the consequentialism I endorse in moral or ethical theory.[7]

As I take it, consequentialism is the doctrine according to which the right option in any choice—the right policy to choose in institutional design, the right alternative to take in personal life—is always the option that promotes the neutral good, maximizing expected neutral value. Neutral or agent-neutral goods are in principle

[6] On the idea of property-desires, see (Pettit 1991b).

[7] Relevant texts include (Pettit and Brennan 1986; Pettit 1984, 1987, 1988, 1989; NJD; Pettit 1991a, 2001b; Pettit and Braithwaite 1993, 1994; CP; R; 1997, 2000b, 2012; Pettit and Scanlon 2000; Pettit and Smith 2000, 2004).

available to be recognized from all points of view and involve properties like pleasure, happiness, peace, justice and so on. They contrast with perspectival or agent-relative goods that are cast in indexical terms and privilege a certain point of view: for example, the good of me or mine, the good of supporting my country, the good of keeping my promises or the good of honoring my debts.

The consequentialist approach can recognize that there are perspectival goods or values, in the sense previously explained. What it maintains, however, is that when it comes to justifying a choice, including any choice that involves acting on a perspectival or agent-relative good, the question is whether that choice is consistent with promoting neutral good or neutral value overall. If it is not, then the choice cannot be justified to others; it remains a self-seeking choice that others have no reason to accept or applaud.

Non-consequentialist theories, in whatever form, hold that justification does not always require identifying a neutral good—a good which is in principle recognizable by all as a common good—that is promoted by the choice. The idea is that sometimes it is enough by way of justification to show that the perspectival good that the choice promotes (or, as it may be said, the perspectival reason for which it is adopted) is a good (or reason) that anyone in your perspective could reasonably have valorized. Like the good of favoring your children or keeping your commitments it is a good (reason) that is concordant with or parallel to goods (reasons) that others are equally liable to recognize in their own perspectives. Justification on the consequentialist model invokes the purportedly common good that the justified choice promotes. Justification on the non-consequentialist alternative may invoke the perspectival good, purportedly concordant with perspectival goods that it is reasonable for others to countenance in their own case, which the justified choice serves.

The more plausible forms of consequentialism agree with non-consequentialists that there ought to be constraints in public life that restrict how agents can make their decisions: legislators should be constrained to make law within the constitution, for example, and judges to adjudicate according to the law. And equally they agree that you ought to establish constraints in your personal life that restrict how you make your personal decisions: you ought not to consider what is best for yourself, or even best for the world, in responding to a friend's requests, for example, in answering a request for information, or in acting on the basis of a promise. Where non-consequentialists take such constraints to be basic, answering to perspectival goods, consequentialists can recognize that the best way of promoting neutral goods may often involve acting under the constraints. Everyone benefits insofar as legislators and judges conform to restrictive briefs, rather than taking the law into their own hands. And your friends and other interlocutors benefit insofar as you treat them generally as friends or as people with certain claims upon you.

Consequentialists and non-consequentialists converge, not only in the recognition of the need for public and personal constraints on decision-making. They generally agree in addition that no such constraints are absolute. There are always going to be emergency cases where it would be right for the holder of a public office to breach an established brief. And there are always going to be situations—perhaps quite perverse and improbable situations—in which it would be right for you to let down a friend, tell a lie or break a promise; it may be that someone's life depends, for example, on your doing so.

Despite these two areas of convergence, consequentialists and non-consequentialists divide on the issue of when it is right to breach public or personal constraints on decision-making. Non-consequentialists appeal to intuition, arguing that there are points where any right may be breached in the name of another right, or even in the name of the public good that honoring that right would jeopardize. Consequentialists have a more straightforward answer, although it rests too on intuitions about the neutral goods that matter. They say that public and personal constraints should be breached when that is essential for the promotion of neutral value overall.

One problem that is often raised for consequentialists is to explain how agents could be alert to those cases where it is right to breach a constraint for the sake of promoting neutral value overall; in particular, to explain how agents can do this without really ceasing to operate as public officials, or as personal friends. The charge is that such agents would always have to keep an eye on what is for the overall neutral good and, indulging what Bernard Williams called 'one thought too many', could not really conform to the constraints we expect to be satisfied. The judge would go through the motions of meeting relevant constraints but would always be doing a calculation about whether this is really for the best. The friend may seem to treat you as a friend, focused only on your good, but would really be strategizing about what the good of the world requires.

The resolution of this problem for which I have long argued is that consequentialism requires agents to commit themselves pretty faithfully to their briefs and their roles, operating on automatic pilot, and that they should rely on external cues—red lights—to deliver the message that a case is exceptional and may require a breach of those constraints. Assuming that such sensitivity to cues can be reliable, the approach supported would not involve agents in the exercise of any counter-productive monitoring or strategizing. They would behave as friends when asked to move an apartment but would recognize the need to think more broadly when asked to move a body. They would behave as honest brokers when asked directions by a passerby but would think again when asked for the whereabouts of his wife by an angry, abusive husband.

With this background in place, I can address the issues raised in the two relevant papers. The authors of 'Two Problems of Value-Monistic Consequentialism' address the consequentialist theory of criminal justice which I defended in joint work with John Braithwaite. According to that theory, the criminal justice system considered as a whole—that is, as a system of criminalization, policing, prosecution, adjudication and sentencing—ought to be designed to further the republican goal of freedom as non-domination; more on this later. The case we make is that the system that this goal supports would be functionally effective, satisfying desiderata of acceptance, stability and satiability, and that it would offer useful guidelines for the design of different aspects of the system, ranging from what should be criminalized to what sentences should be attached to different crimes, to how much discretion should be allowed to the different agents of the system, in particular the judges.

The authors of the paper wonder whether the republican consequentialism endorsed here can support the sorts of rights most of us would want established in the criminal justice system and whether it identifies a sufficiently sharp goal for the system to promote. The comments just made are designed to help me deal with those two challenges.

In response to the first, I say that there is no reason to doubt the capacity of a system that is designed overall for the promotion of a certain good to be able to establish rights for the various agents within it: say, the right of a defendant to a trial by jury. John Rawls argued in his early essay on 'Two Concepts of Rules', that a consequentialist might defend a court system under which the judges are strictly bound to abide by certain rules – and to exercise discretion only to the extent that that is allowed by the rules – on the grounds that such a system does best overall by the preferred goal. I say the same here. The system ought to be able to guard against the corrupt judge or other agent by recourse to the usual measures of correction. And it ought not to motivate the zealous judge or agent to overstep the bounds, say by seeking an exemplary punishment in a given case or by overlooking certain evidence. Such an official ought to be able to recognize that such morally motivated opportunism is extremely hazardous, since discovery would have dire effects, undermining popular confidence in the system.

What of the second complaint, that the goal of promoting freedom as non-domination – or dominion, as Braithwaite and I then called it – is just too vague to be useful? In response I make two points. First, that the goal is meant to guide the designers of the system in the first place and not, as the paper sometimes seems to assume, the agents within the system. And second, Braithwaite and I invoke certain interpretative guidelines or heuristics in order to make this goal more specific: guidelines, for example, on which sentences are likely to serve freedom as non-domination best. It is no objection to those guidelines that they are not deducible formally from the more abstract goal, as the paper suggests; they are designed precisely to meet the original complaint of vagueness.

The authors of the other paper, 'Indirect Consequentialsm and Moral Psychology', give a very nice account of my indirect consequentialism before raising two problems. The first is that even if an agent really goes on automatic pilot in abiding by constraints of friendship or honesty, for example – the paper airs some doubts about how far this is possible – he or she will still retain a consequentialist viewpoint in assessing the performance of others. And so, the suggestion goes, that agent will be schizoid in some measure. I am not moved deeply by this issue. I see no reason why I as an agent who is moved consequentially should not assess others for their friendship or honesty without falling into a schizoid mentality. I might assess others in the ordinary way for how reliable they are as friends or informants. And at the same time I might assess them for whether they are so locked into those roles that they would not be able to break frame even for the sake of avoiding some palpable ill.

The other problem raised by the authors of this paper cuts much more deeply. They take issue with my saying in some texts that when you act in a friendship or honesty role, as consequentialism requires, you go modular, allowing a subsystem in your psychology to take over what you do: this, in the way in which you allow your fingers choose the keys if you are a good typist. They argue, quite rightly, that to act out of friendship or honesty or any such motive is never to go modular in this more or less mindless way. And so I concede that I should not have written as I did. I absolutely agree with them that to act out of friendship or whatever is still to remain in reasoning mode and not to follow a disposition blindly or mechanically. But of course I think that there is nothing in my indirect consequentialism, as outlined above, to prevent me from acknowledging this point. Acting out of friendship or honesty, as I now see it, is not to give up on being guided by reasons. It is just to accept that there is an exclusionary reason in play – provided the red lights do not go on – that blocks you from deliberating in a manner that is inconsistent with being a good friend or an honest broker.

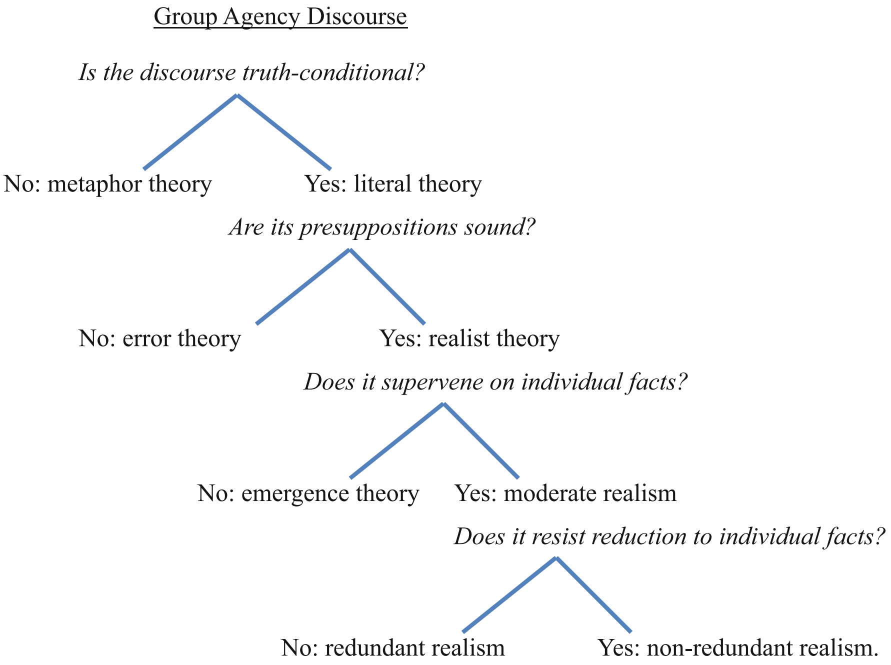
13.4 Group Agency

Two papers address my views about group agency and about related matters in social ontology. One is 'What is the Foundations of Pettit's Non-redundant Realism about Group Agents' and the other is the single-authored paper by David Schweikard, 'Pluralism across Domains'. Before addressing the points raised in the papers, let me set out the main elements in the view I have defended, often in collaboration with Christian List.[8]

In any area of discourse, including discourse that treats groups as agents, there are a series of questions to raise. Applied to talk of group agency the questions generate a tree on the pattern in the attached figure. Some of our talk about group agents, as when we speak of what the markets think or what generation X prefers, is clearly metaphorical or expressive. But other talk, I assume, is quite literal and truth-conditional: for example, talk about the goals or judgments of a corporation or a church or a political party. Does such talk make false presuppositions, as in positing entities that simply do not exist? Surely not: there is nothing obviously misconceived about positing group agents and attributing goals and judgments, as we might attribute them to individuals. In order to give countenance to this talk, do we have to assume that corporate bodies emerge from individual action, as under the impact of a novel force, so that facts about group agents do not supervene on facts about what the individual members think and do? Again, surely not. There is nothing that happens at the corporate level that is not fixed superveniently by the attitudes and actions of members. Finally, is there any reason to think of group agents as agents in their own right, agents with minds of their own, which are not readily reducible to individuals? This is the crucial question and the main thrust of my work, in

[8] The most important publications on, or in the background to, the topic of group agency are: (CM; TF; Pettit 2001a; List and Pettit 2002, 2004, 2005, 2006; Pettit 2003a, 2003b, 2005b, 2006b, 2007d, 2009a, b; Pettit and Schweikard 2006; GA).

particular my work with List, has been to argue that yes, there is a reason to treat group agents as resistant in this way to individualistic reduction.



Group Agency Discourse

Is the discourse truth-conditional?

No: metaphor theory Yes: literal theory

Are its presuppositions sound?

No: error theory Yes: realist theory

Does it supervene on individual facts?

No: emergence theory Yes: moderate realism

Does it resist reduction to individual facts?

No: redundant realism Yes: non-redundant realism.

I assume, as this diagramming of issues indicates, that there are groups that count as agents. By that I mean that groups of individuals often organize themselves so that, first, they form purposes or goals, perhaps once for all, perhaps in an evolving manner; second, they form evidence-based representations or judgments – on-off in character – about the relative importance of those goals, the emerging opportunities for pursuing them, and the best means to adopt in that pursuit; and third, they act – always, of course, through one or more of their members – so as to pursue those goals in a way that makes sense according to those judgments. In short, there are groups such that when things go to form – when conditions are normal – they simulate the performance of individual human beings.

These assumptions, borne out in our experience of bodies like firms and churches, voluntary associations and political parties, make me a realist about group agents. But I recoil from the sort of realism that would posit forces over and beyond anything that is bound to be contributed just by the presence of individual agents: the members who presumably act in their own right as centers of belief and desire, purpose and representation. And so I have to believe that all that group agents think and do, they think and do superveniently on the attitudes and behavior of their individual members. That then takes me to the question as to whether group agents are readily reducible to their members – whether we can identify their dispositions and doings just by looking at the corresponding dispositions and doings of those members.

In order to count as agents, groups have to display a modicum of rationality, at least when things go to form. They have to be evidentially rational in the formation of their judgments and their associated purposes; and they have to be executively rational in acting after the pattern that is mandated by the judgments and purposes they form. Or if they fail to be rational in one or other of these ways, as we individual human beings often fail, then at least they have to be capable of being called to book and they have to be sensitive to such challenges. They have to be able to recognize that it will not do for them to embrace inconsistent judgments or purposes, for example, or to fail to act as the judgments and purposes embraced require. And they have to be able to acknowledge such failures, when they occur, and to rectify the situation or apologize for failures in the past.

The question of whether group agents are not readily reducible – the question of whether they are autonomous agents, as we can put it, agents in their own right – turns on whether the requirement of being collectively rational in this sense is consistent with ready reducibility. List and I argue that it is not.

Let's agree on conditions under which reducibility should be admitted. I assume that if the attitudes of a group on this or that proposition were a majoritarian function of the corresponding attitudes of the full assembly of members, then reducibility would be assured. I assume that were the attitudes on one set of propositions a majoritarian function of the corresponding attitudes of a sub-unit of the group, and the attitudes on another set of propositions a majoritarian function of the corresponding attitudes of another sub-unit, overlapping or not, then reducibility would also be assured. And I assume that reducibility would equally be ensured if the attitudes in some cases were a non-majoritarian function of the attitudes of the appropriate sub-unit, and if each sub-unit worked with a different function, majoritarian or not, when generating group attitudes in its area of specialization. In all these cases, however complex, the group is wholly responsive to members, generating for any proposition, p, a judgment or preference on whether p or not p in systematic response – in a response that fits a determinate function, majoritarian or not – to the corresponding judgments or preferences of members in the appropriate sub-unit. In all of these cases the group's attitudes are formed on a bottom-up basis, with the relevant members exercising local control over what the attitude on any proposition is to be. Those individuals each have to play their roles in one or another sub-unit, or in the assembly of the whole, with the group's attitudes emerging mechanically in response to their efforts.

The reason I believe that group agents are not readily reducible to individuals, and constitute autonomous agents, is that individual responsiveness of any of the kinds illustrated – and the corresponding forms of reducibility – turn out to be in tension with the collective rationality that any well-functioning agent has to display. The discursive dilemma that I identified, generalizing the doctrinal paradox that Lewis Kornhauser and Larry Sagar had noticed in judicial decision-making, shows that if a group is responsive in the straightforward majoritarian fashion, letting the attitudes of the group be determined by majority opinion among its members, then it cannot be robustly rational. Suppose three members, A, B and C, vote on three connected propositions like p, q and p&q. A and B may vote for p, B and C vote for q, and so A and C vote against p&q. Thus if the group is to get its act together, as we say – if it is to prove capable of performing as a rational agent – then on at least one of those propositions it has to endorse as a group a judgment that a majority reject.

The discursive dilemma shows that a group like our A-B-C trio has to choose between being individually responsive to its members in the majoritarian away and being collectively rational in the manner expected of an agent; there is no middle way. But consistently with that being the case, it might be that a group agent could hope to be collectively rational and yet be individually responsive in a different, more complex manner. At this point the impossibility theorems that have been established over the past decade, beginning with the theorem that List and I provided in 2002, become relevant. What they show is that under any of a whole variety of ways of interpreting individual responsiveness, it remains the case that a group agent cannot be robustly rational; it will be forced at one or another point into inconsistency and irrationality. The lesson is that if a group is to prove itself a rational agent, then it has to be prepared to breach the constraint of individual responsiveness. And that means that it has to be prepared to form attitudes, and to update its attitudes in light of new evidence, on a pattern which means that the attitudes it forms may not be a systematic function of the corresponding attitudes – the attitudes on corresponding propositions – of its members.

The upshot can be put as follows. In order for a group to prove reliable as an agent, the members have to be ready to abandon a purely bottom-up way of determining group attitudes in favor of a feedback procedure that involves two steps: first, checking on whether the result of any bottom-up measure generates an attitude that fits with the attitudes otherwise adopted by the group; and, second, revising the attitudes that have bottom-up support in order to ensure consistency. Thus a simple majoritarian assembly might hope to prove itself a workable group agent by following a straw-vote procedure under which the members check after every vote on whether the attitude supported fits with existing attitudes and, if it does not, revise one or other of its attitudes in order to ensure consistency. Thus the group in our earlier example, finding that the vote against p&q leads them into inconsistency, would have to reject the result of that vote, or of one of the earlier votes on p and on q. The members might decide that despite their individual views, they should act as if it is the case that p, that q, and that p&q, when they are acting in the name of the group.

With these points in place, I can turn finally to the issues raised in the two papers related to my theory of group agency. Both the three authors of the first paper, and David Schweikard, draw attention to the fact that in the introduction to the book with Christian List I say that the theory is consistent with 'the methodological individualism of the philosopher Karl Popper, the economist Friedrich Hayek, and many others'. This, as they point out, jars with the fact that elsewhere I reject what I describe as methodological individualism in favor of a methodological or explanatory pluralism: an approach under which it is sometimes appropriate to explain social events by reference to individuals and sometimes to explain them in a fashion that abstracts from individuals. David Schweikard exercises due charity, however, in pointing out that in espousing Popperian individualism I mean only to say that social explanation – and in particular the explanation of the actions of group agents – does not depend on postulating any mysterious forces. Thus the Popperian position I adopt in the book with List is consistent with my rejecting methodological individualism in another sense: in the sense in which it would prescribe always seeking to explain things by reference to lower levels, and in the social case by reference to individuals. As Schweikard points out, my pluralism on the explanatory front fits with the program model that I defend in a number of domains, and not just in the social case.

While the focus is on Schweikard's paper, I should take the opportunity to enter an amendment of my views on a matter he very usefully discusses. He points out that List and I use the program model to explain how it can be the case both that a group agent should be held responsible for something it does via one or more members and that those members should also be held responsible. The group agent, as I put it, is fit to be held responsible in virtue of arranging for the action to be done by this set of members or, if they fail, by that other, substitute set: that is, in virtue of necessitating the action, as we can say. And the members are fit to be held responsible, at least in many cases, insofar as they could have refused to perform the action and forced the group to look to substitutes instead. The amendment I would wish to add now is that a group agent may be fit to be held responsible for something done in its name just by virtue of allowing it, not necessitating it; it can be fit to held responsible in virtue of not having adopted steps to prevent its members from taking the course of action involved.

Back now to the other, multi-authored paper. The authors have one further complaint I should address. This is that my reference to mysterious forces in describing what Popper's methodological individualism – better, ontological individualism – involves is itself mysterious. Since I describe them as psychologically mysterious forces, they say, then I should appeal to contemporary psychology to show why they are mysterious. But here I should clarify. When I say that the theory of group agency defended does not postulate mysterious forces – and so is ontologically individualist – all I mean to say is that it postulates no forces other than those associated with the psychologically unexceptional performance of individual members: that is, with their performance as agents who are moved in regular, intentional fashion by the purposes and representations they form.

One final issue arises in connection with both papers. What does it mean in practice to hold that while group agents involve no mysterious forces in this sense, nonetheless they should be regarded as agents in their own right, not merely as projections of their individual members? I say that it means much the same as it means to say that while individual agents involve no mysterious forces over and beyond what their neurons provide, still they have to treated as agents in their own right; they have to be regarded from within the intentional stance as systems with an integrity of their own.

The point can be made with appeal once again to the program model. We might be able to explain the doings of an individual in neuronal terms and yet miss out on the information provided in an intentional explanation: that the agent would have behaved in just that way, under different neuronal antecedents, provided that he or she still retained the attitudes invoked in the intentional story. The same holds in the group case. We might be able to explain this or that adjustment on the group's part by reference to what certain individuals did and yet fail to see that the group would have acted in the same way, despite differences in the contributions of particular individuals, provided that it retained the attitudes that we would invoke in a group-level, intentional explanation of its behavior.

This is to say that just as the dependence of biology on chemistry and chemistry on physics allows room for useful biological and chemical explanation, so the same is true of group-level explanation in relation to individual-level and of individual-level explanation in relation to neuron-level. But there is something else to add as well. This is that many of the most important human activities are possible only insofar as we adopt the intentional stance, treating individuals in abstraction from neuronal knowledge and treating groups in abstraction from individualistic. I cannot pursue conversational exchange with an individual or group agent – I cannot try out reasons on such an agent or seek out commitments from them – without confining myself to the intentional stance; it is only within such a stance that the notion of reason or commitment makes sense. Not only does the inference to the best explanation argue for treating individual and group agents as autonomous systems, then; so does the indispensability of that treatment for conducting the most important forms of human exchange.

13.5 Republicanism

The final two papers among the commentaries on my work focus on my commitment to republican political theory. They are 'Which Liberalism, Which Republicanism? Constructing Traditions of Political Thought with Pettit' and 'Focusing on the Eyeball Test: A Problematic Testing Device in Philip Pettit's Theory of Justice'. As in the earlier sections, I will set out my overall position with a view to addressing the points raised in these papers. And then I will consider their particular criticisms.[9]

The republican theory with which I identify is primarily distinguished by the conception of freedom as non-domination. I dominate you in a choice to the extent that I have a power of interfering with you which you do not control – I have an arbitrary power of interference, as it used to be said – whether or not I actually exercise that power. You do not enjoy freedom in a dominated choice of this kind because you depend on my will for being able to choose as you will. I am the one in

[9] Relevant works are: (Pettit 1996; R; 1998b, 1999a, 2000a; TF; 2002a, 2003a, 2005a, 2006a, 2007b, c, 2008a, b, d, 2010, 2011, 2013; Lovett and Pettit 2009; Marti and Pettit 2010; GA; OPT).

ultimate charge of what you do; you can act as you wish only insofar as I permit you to act as you wish.

On this account there can be domination without interference, and so a loss of freedom without interference. The paradigm case is that of the slave of the benign master; this lucky slave may get what he or she wants but does so only because that is the master's will. On this account, equally, there can be interference without domination, and so a freedom that survives interference. A striking example is that of the person who allows another to interfere in their life, as when Ulysses allows, indeed orders, his sailors to keep him bound to the mast.

This way of thinking about freedom, as I have argued in many places – building on the work of historians like Quentin Skinner – was the standard conception of free choice in the European tradition from the time of the Roman republic down to the end of the eighteenth century. Writers in that tradition focused, not on freedom in a particular choice, but on freedom in such a range of choice – ideally, a range of choice available to all citizens of a republic – that it makes the person free: it establishes the person as a *liber*, in the Latin usage, or as a 'freeman' in the standard English translation. They all argued that the free person could not be anyone's subject or subordinate or slave, no matter how benevolent the master; in that respect they were social radicals, though the radicalism was confined in the fashion of premodern times to propertied, mainstream males. And they all argued that if the interference of the state was subject to the equally shared control of an active citizenry – if it satisfied republican constraints – then the law imposed by the state would not be dominating and would not deprive citizens of their free status. In that respect, traditional republicans were not just social radicals but political believers: they thought of the law and the state as a precondition of freedom, not as something inimical to it.

The tradition of republican thought begins in classical Rome, finding expression in the work of Polybius, Cicero and Livy. It reappears in medieval and Renaissance Italy where its most outstanding spokesperson is Machiavelli of the *Discourses on Livy*. It ignites the English civil war in the seventeenth century and, becoming reconciled with a constitutional monarchy, shapes English-speaking political thought right through to the end of the eighteenth. Here the great writers are Harrington and Sidney in the seventeenth century and the many supporters of the American revolution in the eighteenth: for example, Price and Priestley in Britain, Jefferson and Adams, Hamilton and Madison, in the new United States.

In all of these writers we find three recurrent themes, though they receive different interpretations and emphases in different hands. First, that freedom is non-domination and that the rationale of the republic is to ensure the free, un-dominated status of its citizens; second, that in order to achieve this goal the republic must be organized under a mixed constitution: a rule of law that gives mutually checking powers to distinct public authorities; and third, that in order for the mixed constitution to work appropriately in pursuit of that goal the citizenry have to enjoy and exercise considerable electoral and contestatory power.

In setting up this picture of the republican tradition of thought, I have generally contrasted it with the new way of thinking about freedom, anticipated in Hobbes, that was introduced in the late eighteenth century by utilitarians like Jeremy Bentham. Under this novel conception – Bentham described it as ‘a kind of discovery’ – freedom merely required non-interference. That conception became the mainstay of the libertarian or classic liberal school of thought that arose early in the nineteenth century. While implicitly accepting that the citizenry might include workers and women as well as propertied males, those libertarians argued that the state did not have to concern itself with domination in circumstances where interference was, as they thought, unlikely: say, in the factory setting where the rational employer allegedly has no economic motive for throwing his weight around, or in the domestic context where the legally more powerful husband can be supposedly expected to honor Christian charity and mercy in dealing with his wife. And equally they argued that since all interference, including the interference of a coercive law, takes away freedom, the cause of freedom as non-interference argues for relying as little as possible on the state. Where republicans – or at least republicans who extended their conception of the citizenry – were social radicals and political believers, these new classical liberals were social conservatives and political skeptics. They thought, in the words of Ronald Regan, that government is the problem, not the solution, and put their faith in the working of a free, unregulated market.

The classical liberals, right-wing in orientation, were challenged in time by a new breed of left-wing liberals such as John Stuart Mill and, in our own period, John Rawls. These figures continue to embrace the conception of freedom as non-interference, losing sight of the older republican ideal, but insist that ideals like equality and welfare belong with the ideal of freedom as non-interference among the goals that the state ought to invest in promoting.

Many contemporary political philosophers, myself among them, identify as neo-republicans, arguing that we should rethink the state as a body that has the job of securing the free status of all its citizens, operating under the discipline of a mixed constitution and a contestatory citizenry. Neo-republicanism is a research program whose aim is to explore how far it is possible to derive appealing institutional designs and policy initiatives on the basis of a concern with freedom as non-domination. I have argued that that ideal can support an attractive, yet feasible vision of the requirements, on the domestic front, of social justice and political democracy and, on the international, of the sovereignty that peoples ought to enjoy in relation to one another, to multi-national forces, and to international agencies.

In sketching the history and outline of republicanism, I have said little or nothing about Rousseau and Kant. I think that both philosophers endorse broadly a notion of freedom as non-domination, sticking with the old tradition. But I hold that Rousseau – and to a lesser extent Kant – was driven in a very different direction from that of the Italian-Atlantic republicanism I have described by rejecting the ideal of the mixed constitution and, with it, the associated ideal of the contestatory citizenry. Under the influence of Bodin and Hobbes, Rousseau assumed that no state could function properly unless there was a unitary power in charge – a single sovereign. He argued that this sovereign had to be an assembly of all the citizens, operating – impossibly, as we know from the discursive dilemma – under a regime of majority voting. And so he gave citizens a participatory role in that assembly, denying them the right to contest what the assembly in its wisdom – and ideally, if improbably, in expression of the general will – decided. Thus he ends up defending a communitarian or populist picture in which citizens are comprehensively subordinated to the collectivity in order to be made independent of one another: 'each, by giving himself to all, gives himself to no one'.

The authors of 'Which Liberalism, Which Republicanism?' take issue on a number of points with the story, in particular the historical story, that I tell about republicanism, liberalism and communitarianism. In a more philosophical vein, they complain that I suggest that liberalism did not have the resources to criticize slavery. I reject the complaint, since I emphasize in my 1997 book on Republicanism, on which they almost entirely rely, that left-wing liberalism often converges on matters of justice – and certainly in opposition to slavery – with the republican theory I support. Equally, while I think that right-wing liberalism is soft on subordination – to this day, libertarians argue for the right of an employer to fire at will – I nowhere suggest that with their extended conception of the citizenry, they would tolerate slavery.

Most of the complaints in the paper are historical in character but my sense is that, like this first complaint, they stem from misunderstanding. Thus I do not think that Hobbes is a liberal, only that he anticipated the conception of freedom later adopted by liberals; he was not a liberal because freedom, even in that sense, was not an ideal that guided his thought and policy. And neither do I suggest that Filmer is a liberal when I point out that he shows signs of going along with the Hobbesian view of freedom. Nor finally do I break ranks with my earlier, freedom-centered characterization of republicanism when I say that John Locke supports some republican ideas. After all, Locke is famous for thinking of liberty as something that a well-ordered law does not take away: 'that ill deserves the name of confinement that hedges us in only from bogs and precipes'.

Another complaint made in this paper is that I do not explain why I give prominence to contestation, downplaying participation. The ideal of participation – that is, the ideal of universal citizen participation in the sovereign, law-making body – belongs with the Rousseauvian version of republicanism rather than with the older tradition, which emphasized representation and contestation. I think that such universal participation is infeasible in practice and that hailing it in theory can have two highly undesirable effects. First, it can easily take the ideals of the mixed constitution and the contestatory citizenry out of the picture, even delegitimize them, as in Rousseau himself. And second, it can easily encourage a departure from the ideal of freedom as non-domination and an embrace of a vulgar ideal that Rousseau rejected but many Rousseauvians embrace: this is the conception of freedom as consisting in the right and exercise of a participatory role in a self-determining community.

The authors of this commentary conclude by voicing the concern that I am misled by the desire for a universal ideal, relevant on all places and times. My response to that criticism is that I think of the ideal as structurally but not substantively universal. Consider Amartya Sen's argument that we should not conceive of poverty either as living below the subsistence level or as living in the bottom 5 % of your society; we should not conceive of it in substantively universal terms or in wholly relative terms. Rather we should think of poverty as lacking the resources and capabilities that are necessary for basic functioning within your local society. That is a structurally universal conception of poverty and, by implication, of non-poverty. To escape poverty in any society is to have the capabilities of basic functioning there, so that this is a universal ideal that applies everywhere. But the ideal requires quite different resources in different societies: basic functioning requires very different capabilities in an advanced urban context from what it requires in a simple agricultural life. And that is to say that the ideal is universal in structure only; substantively it differs from society to society in what it requires.

As it is with poverty and non-poverty in Sen's vision so, on my account, it is with freedom: that is, with the freedom of the person, conceived in terms of non-domination. I say that to enjoy freedom in relation to others in your society what is required is that you have access under law and norm to such powers and protections that you can look all others in the eye, without reason for fear or deference: fear of their interference or deference to their power of interfering. Or at least you can do so unless you count, by local standards, as excessively timid, even paranoid. The idea is that society should give you sufficient support to enable you, if you are not lacking in backbone, to deal with others as equals, expecting and commanding their respect.

As non-poverty requires different resources in different cultures and societies, so freedom as non-domination is certain under this test of adequacy to require different powers and protections in different societies. If you live in a small agricultural society, where almost everybody knows everybody then it may be enough for being able to look others in the eye that there are norms that promise ignominy for anyone who tries to throw his or her weight around, whether domestically or in public. If you live in a contemporary urban society, where most of the people you meet are strangers, this is hardly going to be enough to provide you with that capacity. It will also be essential that there are laws against interference and an adequate public system of monitoring and defense. And so on for other possible sources of variation.

But while it delivers me from the charge of excess universalism, this resort to the idea of the eyeball test exposes me to the main charge of the second commentary, 'Focusing on the Eyeball Test'. The authors of this paper are not impressed by my resort to this test, charging me in the first two of three criticisms with a certain circularity. They allege that to be able to pass the eyeball test amounts to nothing more, in effect, than escaping domination in the range of the basic liberties: that is, the maximal choices, to be identified in a culturally appropriate way under any suitable rule of law, that each can exercise and enjoy at the same time as others. And so they charge me with a corresponding circularity. To understand what non-domination requires, so they suggest, you need to understand how the test works; to understand how the test works you need to understand what non-domination requires.

My response to this charge, of course, is that the eyeball test presupposes local standards of timidity and that these will naturally facilitate its interpretation in this or that context. I assume that the members of any society and culture will be able to tell – without independent access to the abstract concept of non-domination – just when it is possible by those standards to look another in the eye, and when it is not.

I think of the knowledge required as a participant, practical form of knowledge, essential for purposes of local survival and success, which no competent member is going to lack. It constitutes a sort of know-how that enables people to judge about the expectations of others in any interaction and to tell what they can and cannot say and what overtures they should and should not contemplate.

But if the contextualized, locally interpretable character of the eyeball test means that I can avoid the circularity objection, there is still trouble in store. For the authors say that this reference to local standards deprives the ideal of non-domination of the universal status I would ascribe to it and that it would make it next to impossible to compare two societies for how well they do by that ideal. I answer that the ideal can retain the status of a structurally universal ideal, which is all that I would want to claim for it. I concede that it may make it more difficult to compare two societies for how well they serve such a structural ideal, as Sen's characterization of poverty makes it more difficult to compare two societies for how well they do in avoiding poverty. But it will still be possible to make fairly reliable comparisons on an intuitive basis. And it will certainly be possible to identify gross differences in how different societies score in the freedom stakes.

Why, finally, do I give importance to the eyeball test? I think that the addressees of political theory ought to include ordinary people, not just fellow academics and not just the functionaries of government. And this being so, I think it is important for any political philosophy to be able to engage people's intuitions, giving them a heuristic for determining how far they should be satisfied or unsatisfied with the status quo. I see the eyeball test as serving this function with the republican theory of justice, as I see the tough-luck test described in *On the People's Terms* as serving that function with the republican theory of democracy.

I am encouraged in that belief by the fact that when he introduced the law governing same-sex marriage, President Zapatero of Spain invoked precisely this test in support of that avowedly republican policy. Challenging his fellow parliamentarians, he asked whether any of them were ready to look their homosexual friends in the eye and say: I deny you the right to have your intimate relationships given the civic status that we heterosexuals can claim. That was a good question and it reverberated with effect among ordinary Spanish people, eventually generating almost 65 % support for a change in the law.

This is a fitting point at which to stop. In conclusion I would like to offer my warm thanks to all the commentators for the challenging questions they raised in different areas of my work. I am deeply honored by the attention they have given that work and I hope that my replies will communicate the depth of my appreciation and admiration.

References

Jackson, F., and P. Pettit. 2002. Response-dependence without Tears. *Philosophical Issues* (supp. to Nous) 12: 97–117.

Jackson, F., P. Pettit, and M. Smith. 2004. *Mind, morality, and explanation: Selected collaborations*. Oxford: Oxford University Press.

Jaegwon, K. 1984. Ephiphenomenal and supervenient causation. *Midwest Studies in Philosophy* 9: 257–270.

Kim, J. 1998. *Mind in a physical world: An essay on the mind-body problem and mental causation*. Cambridge: MIT Press.

Lewis, D. 1986. *Philosophical papers Vol 2*. Oxford: Oxford University Press.

List, C., and P. Pettit. 2002. Aggregating sets of judgments: An impossibility result. *Economics and Philosophy* 18: 89–110.

List, C., and P. Pettit. 2004. Aggregating sets of judgments: Two impossibility results compared. *Synthese* 140: 207–235.

List, C., and P. Pettit. 2005. On the many as one. *Philosophy and Public Affairs* 33: 377–390.

List, C., and P. Pettit. 2006. Group agency and supervenience. *Southern Journal of Philosophy* 44(Spindel Suppl): 85–105.

Lovett, F., and P. Pettit. 2009. Neo-republicanism: A normative and institutional research program. *Annual Review of Political Science* 12: 18–29.

Marti, J.L., and P. Pettit. 2010. *A political philosophy in public life: Civic republicanism in Zapatero's Spain*. Princeton: Princeton University Press.

Pettit, P. 1984. Satisficing consequentialism. *Proceedings of the Aristotelian Society* 58: 165–176.

Pettit, P. 1987. Universalizability without utilitarianism. *Mind* 96: 74–82.

Pettit, P. 1988. The consequentialist can recognise rights. *Philosophical Quarterly* 35: 537–551.

Pettit, P. 1989. Consequentialism and respect for persons. *Ethics* 99: 116–126.

Pettit, P. 1990a. The reality of rule-following. *Mind* 99: 1–21; reprinted in P. Pettit 2002 Rules, Reasons, and Norms, Oxford, Oxford University Press.

Pettit, P. 1990b. Affirming the reality of rule-following. *Mind* 99: 433–439.

Pettit, P. 1991a. Consequentialism. In *A companion to ethics*, ed. P. Singer. Oxford: Blackwell.

Pettit, P. 1991b. Decision theory and folk psychology. In *Essays in the foundations of decision theory*, ed. M. Bacharach and S. Hurley. Oxford: Blackwell; reprinted in P. Pettit 2002 Rules, Reasons, and Norms, Oxford, Oxford University Press.

Pettit, P. 1991c. Realism and response-dependence. *Mind* 100: 587–626; reprinted in P. Pettit 2002 Rules, Reasons, and Norms, Oxford, Oxford University Press.

Pettit, P. 1996. Freedom and antipower. *Ethics* 106: 576–604.

Pettit, P. 1997. Republican theory and criminal punishment. *Utilitas* 9: 59–79.

Pettit, P. 1998a. Noumenalism and response-dependence. *Monist* 81: 112–32; reprinted in P. Pettit 2002 Rules, Reasons, and Norms, Oxford, Oxford University Press.

Pettit, P. 1998b. Reworking sandel's republicanism. *Journal of Philosophy* 95: 73–96.

Pettit, P. 1998c. Terms, things and response-dependence. *European Review of Philosophy* 3: 61–72.

Pettit, P. 1999a. Republican liberty, contestatory democracy. In *Democracy's value*, ed. C. Hacker-Cordon and I. Shapiro. Cambridge: Cambridge University Press.

Pettit, P. 1999b. A theory of normal and ideal conditions. *Philosophical Studies* 96: 21–44; reprinted in P. Pettit 2002 Rules, Reasons, and Norms, Oxford, Oxford University Press.

Pettit, P. 2000a. Democracy, electoral and contestatory. *Nomos* 42: 105–144.

Pettit, P. 2000b. Non-consequentialism and universalizability. *Philosophical Quarterly* 50: 175–190.

Pettit, P. 2001a. Deliberative democracy and the discursive dilemma. *Philosophical Issues (supp to Nous)* 11: 268–299.

Pettit, P. 2001b. In *Non-consequentialism and political philosophy*, ed. R. Nozick and D. Schidmtz. Cambridge: Cambridge University Press.

Pettit, P. 2002a. Keeping republican freedom simple: On a difference with Quentin Skinner. *Political Theory* 30: 339–356.

Pettit, P. 2002b. *Rules, reasons, and norms: Selected essays*. Oxford: Oxford University Press.

Pettit, P. 2003a. Deliberative democracy, the discursive dilemma, and republican theory. In *Philosophy, politics and society Vol 7: Debating deliberative democracy*, ed. J. Fishkin and P. Laslett, 138–162. Cambridge: Cambridge University Press.

Pettit, P. 2003b. Groups with minds of their own. In *Socializing metaphysics*, ed. F. Schmitt. New York: Rowan and Littlefield.

Pettit, P. 2003c. Looks as powers. *Philosophical Issues* (supp. to Nous) 13: 221–252.

Pettit, P. 2004. Motion blindness and the knowledge argument. In *The knowledge argument*, ed. P. Ludlow, Y. Nagasawa, and D. Stoljar. Cambridge: MIT Press.

Pettit, P. 2005a. The domination complaint. *Nomos* 86: 87–117.

Pettit, P. 2005b. Rawls's political ontology. *Politics, Philosophy and Economics* 4: 157–174.

Pettit, P. 2006a. The determinacy of republican policy: A reply to McMahon. *Philosophy and Public Affairs* 34: 275–283.

Pettit, P. 2006b. When to defer to a majority—And when not. *Analysis* 66.

Pettit, P. 2007a. Consciousness and the frustrations of physicalism. In *Minds, worlds and conditionals: Themes from the philosophy of Frank Jackson*, ed. I. Ravenscroft. Oxford: Oxford University Press.

Pettit, P. 2007b. Free persons and free choices. *History of Political Thought* 28: 709–718.

Pettit, P. 2007c. A republican right to basic income? *Basic Income Studies* 2(2): Art 10.

Pettit, P. 2007d. Responsibility incorporated. *Ethics* 117: 171–201.

Pettit, P. 2008a. The basic liberties. In *Essays on H.L.A.Hart*, ed. M. Kramer, 201–224. Oxford: Oxford University Press.

Pettit, P. 2008b. Freedom and probability: A comment on Goodin and Jackson. *Philosophy and Public Affairs* 36: 206–220.

Pettit, P. 2008c. Physicalism without Pop-out. In *Naturalistic analysis*, ed. D. Braddon-Mitchell and R. Nola. Cambridge: MIT Press.

Pettit, P. 2008d. Republican liberty: Three axioms, four theorems. In *Republicanism and political theory*, ed. C. Laborde and J. Manor. Oxford: Blackwells.

Pettit, P. 2009a. Corporate responsibility revisited. *Rechtsfilosofie & Rechtstheorie* 38(Special issue: Philip Pettit and the incorporation of responsibility): 159–76.

Pettit, P. 2009b. The reality of group agents. In *Philosophy of the social sciences: Philosophical theory and scientific practice*, ed. C. Mantzavinos, 67–91. Cambridge: Cambridge University Press.

Pettit, P. 2010. A republican law of peoples. *European Journal of Political Theory, Special Issue on 'Republicanism and International Relations'* 9: 70–94.

Pettit, P. 2011. The instability of freedom as non-interference: The case of Isaiah Berlin. *Ethics* 121: 693–716.

Pettit, P. 2012. The inescapability of consequentialism. In *Luck, value and commitment: Themes from the ethics of Bernard Williams*, ed. U. Heuer and G. Lang. Oxford: Oxford University Press.

Pettit, P. 2013. Two republican traditions. In *Republican democracy: Liberty, law and politics*, ed. A. Niederbeger and P. Schink. Edinburgh: Edinburgh University Press.

Pettit, P., and J. Braithwaite. 1993. Not just deserts, even in sentencing. *Current Issues in Criminal Justice* 4: 225–239.

Pettit, P., and J. Braithwaite. 1994. The three Rs of republican sentencing. *Current Issues in Criminal Justice* 5: 318–325.

Pettit, P., and G. Brennan. 1986. Restrictive consequentialism. *Australasian Journal of Philosophy* 64: 438–455.

Pettit, P., and M. Smith. 1990. Backgrounding desire. *Philosophical Review* 99: 565–92; reprinted in F. Jackson, P. Pettit and M. Smith, 2004, Mind, Morality and Explanation, Oxford, Oxford University Press.

Pettit, P., and M. Smith. 1996. Freedom in belief and desire. *Journal of Philosophy* 93: 429–49; reprinted in F. Jackson, P. Pettit and M. Smith, 2004, Mind, Morality and Explanation, Oxford, Oxford University Press.

Pettit, P., and T.M. Scanlon. 2000. Consequentialism and contractualism. *Theoria* 66: 228–245.

Pettit, P., and D. Schweikard. 2006. Joint action and group agency. *Philosophy of the Social Sciences* 36: 18–39.

Pettit, P., and M. Smith. 2000. Global consequentialism. In *Morality, rules and consequences*, ed. B. Hooker, E. Mason, and D.E. Miller. Edinburgh: Edinburgh University Press.

Pettit, P., and M. Smith. 2004. The truth in deontology. In *Reason and value: Themes from the moral philosophy of Joseph Raz*, ed. R.J. Wallace, Pettit Philip, S. Scheffler, and M. Smith, 153–175. Oxford: Oxford University Press.