Studies in the Ethics
of Coordination and
Climate Change
Vol. 1

Studies in the Ethics of Coordination and Climate Change Vol. 1

Editors: Tim Campbell
& Olle Torpman

Institute for Futures Studies
Working Papers 2024:1–9
Stockholm, 2024

The Institute for Futures Studies is an independent research foundation financed by contributions from the Swedish Government and through external research grants. The institute conducts interdisciplinary research on future issues and acts as a forum for a public debate on the future through publications, seminars and conferences.

This is a collection of preprints from the research project Ethics of Coordination, financed by Riksbankens Jubileumsfond.

Cover: Tove Salomonsson
Cover image: Unsplash/Lance Anderson
Print: Elanders, Stockholm, 2024
Distribution: The Institute for Futures Studies

Contents

Preface

The Ethics of Coordination project is about to complete its second year. It is hosted by the Institute for Futures Studies in Stockholm, and is generously financed by Riksbankens Jubileumsfond (the Swedish Foundation for Humanities and Social Sciences). The project is led by PI Krister Bykvist. It aims to formulate a new approach to addressing collective harm problems, an ethics of coordination that incorporates both individual and collective duties, and to apply this approach to climate change.

The project has three parts. Part 1 gives a theoretical justification for duties to coordinate. Part 2 focuses on direct individual duties regarding climate change. Part 3 explores duties of coordination in the context of climate change.

The three parts are represented in this first volume of the program's preprint series, consisting of nine papers in total. The papers by Andrić, Blomberg and Petersson, Bykvist and Klint Jensen, and Schwenkenbecher are concerned with foundational issues regarding duties to coordinate and thus belong to Part 1. Campbell and Kaczmarek, Hormio, and Torpman address issues related to direct individual climate duties (Part 2). The focus of the papers by Nefsky and Tenenbaum, and Budolfson deal with topics that are relevant for climate duties to coordinate (Part 3).

In this volume, the papers are presented in alphabetic order of their authors' names. The first paper, by Vuko Andrić, examines the idea that unstructured groups, not just their individual members, can have moral duties and explores how these collective duties might be based on individual duties. It raises questions about how individual duties relate to collective duties, highlighting potential issues of circularity in defining these duties. Andrić suggests that individual duties grounding collective moral duties should be seen as rational duties of moral agents, and argues that whether these duties are perspective-dependent should be left to broader ethical theories, not specific accounts of collective duties.

In the second paper, Olle Blomberg and Björn Petersson address the issue of collective moral obligations, where a group can violate a duty without each individual member being at fault, which complicates moral blame. It critiques the view that moral blame should always evoke guilt and punishment for individuals, arguing that this perspective can seem unfair when applied to group members. The authors propose that individuals can justifiably feel guilt for their group's actions through strong identification with the group, thus reconciling the idea of collective obliga-tions with the concept of moral blame.

The third paper, by Mark Budolfson, examines the issue of collective self-defeat-

ingness. As shown by prisoner's dilemmas, self-interest is sometimes directly collectively self-defeating in that everyone ends up worse off following the self-interest strategy rather than some other strategy. However, it has been claimed by Derek Parfit and others that morality, rationality and other interesting forms of normativity are never directly collectively self-defeating. Using examples, Budolf-son argues against these theorists' claims.

In the volume's fourth paper, Krister Bykvist and Klint Jensen examine so-called moral Hi-Lo Cases.They argue that these are not just abstract hypothetical cases but have analogues in real-life cases, including some climate change cases. Furthermore, the authors show that the common appeal to a cooperative stance risks changing the topic by describing a different kind of situation. In the changed situation, the appeal to actual influence rather than mere possible influence on other agents is necessary but even then the cooperative stance may not be manda-tory if it assumes realistic costs.

The fifth paper, by Tim Campbell and Patrick Kaczmarek, examines the normative ramifications of the finding that creating a new person produces more CO_2 emissions than many other lifestyle choices. Climate Anti-Natalism claims that it is often wrong to conceive a new person because of their CO_2 emissions. Campbell and Kaczmarek identify a harm-avoidance principle underlying arguments for Climate Anti-Natalism but argue that this principle has implausible implications.

In the sixth paper, Säde Hormio starts with the observation that moral theories often struggle to justify why individuals should contribute to collective outcomes when their individual impact seems negligible. The focus should shift from isolated acts to patterns of behavior over time, as most real-life collective outcomes are cumulative. The paper argues that the key issue is not individual actions but maintaining coherence between one's values and contributions, as failing to act consistently with one's values in collective settings undermines moral integrity.

In the seventh paper, Julia Nefsky and Sergio Tenenbaum critique Peter Singer's argument that spending money on personal pleasures is morally wrong, likening it to not saving a drowning child. While many responses challenge Singer's principles, they fail to effectively counter his Pond Analogy, which strongly supports his conclusion. The authors argue that the real challenge lies in the Pond Analogy itself, posing a fundamental question for our understanding of morality, and they conclude by outlining their response to this challenge.

The volume's eighth paper, by Anne Schwenkenbecher, is concerned with situations in which moral agents face collective-action problems where individual actions seem ineffective but collective effort can create change. Moral agents use two types of reasoning: I-mode, focused on individual actions and best responses, and We-mode, which prioritizes collective action as part of a group. While both approaches can

sometimes lead to the same outcome, especially when each person's contribution is crucial, I-mode reasoning often falls short in broader cases like voting or reducing carbon footprints, raising questions about the role of We-mode reasoning in guiding moral obligation

In the ninth and final paper, Olle Torpman explores the responsibility-based moral reasons an individual may have to act even when their actions seem to make no difference to the outcome. Torpman distinguishes between prospective responsibility (acting to contribute to good or to avoid harm) and retrospective responsibility (acting to avoid blame for contributing to harm), and argues that reasons for action can be based on both types of responsibility. He moreover argues that such reasons can justify acting in collective impact cases. This framework suggests individuals have moral reasons to act in certain ways to align with their responsibilities, even in such cases.

We are pleased to be able to share this new work from the Ethics of Coordination project. The authors of the papers would greatly appreciate any comments, questions, and objections that you wish to share with them. Contact information is found on the front page of each paper. We would also like to thank Daniel Ramöller and Erika Karlsson for their help in the editorial process of the work with this volume.

Tim Campbell & Olle Torpman
Editors

Vuko Andrić[1]

How Might Collective Duties be Grounded in Individual Duties?[2]

Some philosophers hold that unstructured groups themselves, as opposed to the members of these groups, can have moral duties. There are different accounts of how such collective duties might be grounded in facts about individual duties of the group members. In this paper, I highlight and discuss some questions for these accounts that seem to warrant more exploration than they have received so far. First, if there is a collective duty to φ that is grounded in individual duties, how does φ-ing feature in the individual duties? The accounts that ground a collective duty to φ in individual duties specify these individual duties with reference to φ-ing. But if a collective duty to φ is grounded in individual duties, then, on pain of circularity, the individual duties cannot be specified in terms of a collective duty to φ. Second, are the individual duties that ground collective moral duties themselves also moral duties? Or are the individual duties, rather, rational duties? I will suggest that the individual duties should be classified neither as purely moral nor as purely rational, but rather as rational duties of moral agents. Finally, are the grounding individual duties perspective-dependent, i.e., do they depend on the epistemic situation of the members, as several philosophers have suggested? I argue that accounts of collective obligations should not commit themselves to an answer to this question, but rather leave the question to general ethical theories that do not focus on contexts of collective duties.

[1] Institute for Futures Studies & Linköping University, vuko.andric@liu.se.

[2] Funding from Riksbankens Jubileumsfond (grant number P22-0662) is gratefully acknowledged.

1. Introduction

Can groups have moral duties? Attempts to answer this question are well advised to distinguish between two kinds of groups. On the one hand, there are structured groups, such as companies and universities. Having their own decision procedures, structured groups seem to resemble individual persons in important ways, which suggests, at least initially, an affirmative answer (e.g., French, 1979; List & Pettit, 2011). On the other hand, there are unstructured groups, such as married couples, the passengers on a bus, customers in a supermarket, voters, and consumers. Such groups lack decision procedures, at least the kind of formal, easily recognizable and well established decision procedures that characterize structured groups. This paper focuses on unstructured groups. Can unstructured groups have moral duties that go above and beyond the individual moral duties of their members? Can, for example, the group of all consumers itself have a moral duty to avoid consumption that leads to climate change and the harms associated with it?

There are different ways of arguing for the claim that unstructured groups can have moral duties. Two types of accounts can be distinguished for the purposes of this paper based on whether the account grounds the collective moral duties of a group in facts about the individual duties of the group members. This paper deals exclusively with accounts that involve such a grounding suggestion, leaving aside accounts that argue for the existence of duties of unstructured groups without grounding them in individual duties (e.g., Jackson, 1987; Rosenqvist, 2019; Tännsjö, 2007; Wringe, 2010).

Following Gunnar Björnsson (2020: 132–3, footnotes omitted), we can distinguish four accounts that ground collective duties in individual duties:

> It has […] been suggested that a group's obligation to φ is grounded in the fact that the group would φ, or would be sufficiently likely to φ, if members discharged their individual obligations to:
>
> > *take steps to collectivize:* to transform the group into a group agent that has its own obligation to φ (Collins 2013; cf. Hindriks 2019; Isaacs 2011: 144–54 on "putative obligations");
> >
> > *we-reason:* to identify φ-ing as the optimal solution to a problem that group members cannot solve individually and to deduce their own individual actions based on this (Schwenkenbecher 2018; 2019);
> >
> > *be* prepared to do their part in φ-ing should they be sufficiently certain that others would as well (Aas 2015); or

> *care* to the right extent about what is morally at stake, in the sense of being disposed to (i) pick up information about what reactions and actions tend to promote what is morally important and (ii) be moved by such information when opportunity arises (Björnsson 2014, Forthcoming).[3]

All these accounts, I think, raise some questions that, to the best of my knowledge, have not received much attention yet and that I will highlight in the next three sections. Section 2 examines the moral status the group's φ-ing has when it features in the individual duties that ground the group's duty to φ. Section 3 discusses if the grounding individual duties are moral or rational in nature. Section 4 deals with the question of whether the grounding individual duties depend on the perspective of the individuals that have these duties.

2. The Moral Status of Obligatory Collective Behavior in The Grounding Individual Duties

The collective duty to φ, on the proposals under discussion, is grounded in individual duties of group members to do something or to be a certain way. Moreover, this doing or being is specified in the proposals with regard to the group's φ-ing. The problem is that the group's φ-ing cannot be classified as obligatory in the grounding individual duties. That would be circular. But what, then, *is* the moral status of φ-ing as this collective behavior features in the specifications of the grounding individual duties?

Suggestions are easy to come by, as some are involved in Björnsson's very definitions of the accounts quoted earlier. According to the collectivization account, a group's duty to φ is grounded in individual duties to *transform* the group into a group agent that has its own *obligation* to φ. This suggests understanding the group's φ-ing as *conditionally obligatory*. According to the we-reasoning account, a group's duty to φ is grounded in individual duties to identify φ-ing as the *optimal* solution to a problem that group members cannot solve individually and to deduce their own individual actions based on this. This suggests understanding the group's φ-ing as optimal, i.e., in terms of value. According to the caring account, a group's duty to φ is grounded in individual duties to be disposed to (i) pick up information about what reactions and actions tend to promote what is *morally important* and (ii) be moved by such information when opportunity arises. This suggests understanding the group's φ-ing as *morally important*.

Let us consider the suggestions in reverse order, starting with the idea to characterize the group's φ-ing as *morally important*. A problem suggesting itself is that not every

[3] Björnsson, like other participants in the debate, speaks of collective *obligations*. I use the concepts *duty* and *obligation* interchangeably.

morally important joint action should come out as all-things-considered obligatory. For example, it might be morally important for a group to save people from starvation but more morally important not to achieve this by fraud. The distinction between pro tanto and all-things-considered duties suggests that the group's φ-ing, insofar as it features in grounding individual duties, should better be characterized in terms of being *most morally important.*

However, this characterization is not yet satisfactory either. For it remains unclear how saying that a group's φ-ing would be *most morally important* is different from saying that the group's φ-ing is, or would be, morally obligatory. But if this is actually meant, then we should better be clear about this.

Saying that the group's φ-ing *is* obligatory is different from saying that it *would be* obligatory. If the group's φ-ing is characterized in the individual duties as (actually) obligatory, we are thrown back to the circularity problem elaborated at the outset of this section. If instead calling the group's φ-ing *most morally important* is meant to express that the group's φ-ing *would* be obligatory, this suggestion is indistinguishable from stating a conditional duty of the group to φ, a suggestion that we shall consider below.

Neither interpretation is helpful. But perhaps something else is meant, and characterizing the group's φ-ing as *most morally important* is neither to be understood in terms of actual duty nor in terms of conditional duty. But then it remains unclear how we *should* understand the characterization.

Let us then turn to the suggestion of understanding the group's φ-ing in value terms. Should the group's φ-ing feature as *optimal* in individual duties? The problems facing this suggestion are familiar from the literature on consequentializing (Portmore, 2022). In what follows, I will highlight some of the problems. My point, however, is not that the problems are insurmountable – perhaps they can eventually be solved. Rather, I suggest that it would be problematic to characterize the group's φ-ing as *optimal* in the description of the underlying individual duties because this suggestion invites problems associated with consequentializing, problems that it would be better to avoid (until it has been demonstrated that the problems can be solved, which I take it has not been achieved yet).

First, how does this suggestion deal with constellations in which the group's φ-ing would be *optimal* yet not obligatory? An account of collective duties will probably want to leave conceptual space for *supererogatory* collective actions. But it is unclear how the suggestion can achieve this.

Second, certain moral theories, like rule-consequentialism, have built into them assumptions about the relation between duties and values that cause problems for the suggestion. In particular, rule consequentialists might want to say that a group has (hasn't) a duty to φ only if it would (wouldn't) have the best consequences if groups

in relevantly similar circumstances φ-ed. But this doesn't mean that this group would (not) bring about the best consequences by φ-ing on a particular occasion.

Third, Kantian theories are often characterized as assuming that the right is prior to the good. This does not sit well with the suggestion to characterize the group's φ-ing as optimal and then deduce, via individual duties, that the group has a moral duty to φ.

These problems appear serious. However, let us consider one natural modification of the suggestion (if only to show that it is problematic too). Instead of characterizing the group's φ-ing as *optimal*, we can say that it would be deontically optimal. This is modelled on the suggestion of some consequentializers (e.g., Zimmerman, 1996).

This modified suggestion still raises objections, as two examples shall illustrate. First, "deontic value" is a term of art and as such not connected to everyday moral talk and thought. Therefore, the suggestion has a hard time getting support from ordinary intuitions (consider the parallel problem for the concept of agent-relative value discussed in Schroeder, 2007). This will indirectly affect the overall plausibility of an account of collective moral duties.

Second, while the modified proposal might allow us to accommodate collective supererogation, it remains doubtful how referring to deontic value as something that partly *grounds* moral duties coheres with the explanations of moral duties suggested by moral theories. Does it really fit with Kantianism, say, if we do not say that a group's φ-ing is obligatory because it is the only way of acting in accordance with the Categorical Imperative but instead that the group's φ-ing is obligatory because it is deontically optimal, and then add that it's deontically optimal because it is the only way of acting in accordance with the Categorical Imperative?

Having dealt with the "moral importance" and "optimality" suggestions, let us finally consider the suggestion that the group's φ-ing features as conditionally obligatory in the grounding individual duties. The problems I see with this suggestion concern the grounds of the individual duties. While it is easy to see how individual duties with regard to a group's φ-ing could be based on the fact that the group's φ-ing would be optimal (after all, it is a natural thought that individuals have reasons to promote good states of affairs), the same is not true if the group's φ-ing is merely conditionally obligatory. On the collectivization account, it would be natural to wonder why individuals should take steps to turn a merely conditional duty into an actual duty. What speaks in favor of taking these steps from a moral point of view? (An analogy: By making a promise I can bring it about that I have an obligation, namely to keep that promise. But this do not by itself suggest that I should make a promise.)

The same kind of question suggests itself on the other accounts if they are combined with the suggestion that the group's φ-ing features as conditionally obligatory in the respective individual duties. Why should one care about a merely conditional

collective duty? Why should one we-reason about merely conditional duties? Why should one be prepared to play one's role in collective behavior that is merely conditionally obligatory?

To summarize, according to the accounts of collective duties on offer, a group's duty to φ is grounded in individual duties of the group members, whereby the individual duties concern individual behaviors or personal characteristics that are specified with regard to the group's φ-ing. In this specification, on pain of circularity, the group's φ-ing cannot feature as being a collective duty. But what then is the moral status of the group's φ-ing insofar as it features in the grounding individual duties? I have discussed three suggestions, which classify the group's φ-ing as morally important, optimal, or conditionally obligatory, respectively. I found all three suggestions, including some modifications that came to mind, wanting. The result is that I am not aware of any convincing answer to the question of what moral status the group's φ-ing has insofar as it features in the grounding individual duties.

3. Are the Individual Duties Moral or Rational?

We have considered four accounts of grounding collective moral duties in individual duties: the collectivization, we-reasoning, preparedness, and caring proposals. The question I want to consider next concerns the nature of the grounding individual duties. Are the individual duties themselves moral duties, or are they more plausibly categorized as rational duties? The question is relevant if we try to formulate a comprehensive theory of collective duties, because moral theories do not necessarily coincide with theories of rationality and we need to know how to categorize the grounding individual duties in order to decide how best to account for them, with a moral theory or rather with a theory of rationality.

But how can we approach the question, how can we find out if the grounding individual duties are themselves moral duties or not? My approach will be based on a desideratum for accounts of collective moral duties. The desideratum is that an account of collective moral duties should be compatible with as many (*prima facie* plausible) ethical and meta-ethical positions as possible. This desideratum, I shall argue, speaks in favor of the view that the grounding individual duties are neither purely moral nor purely rational in nature.

Let us begin with the natural view that, since the collective duties we are concerned with are moral, the grounding individual duties are moral as well. The problem with this view is that it carries controversial commitments when combined with the accounts of collective duties under consideration. The desideratum that an account of collective moral duties should be compatible with as many ethical and meta-ethical positions as possible thus speaks against this view.

The problem is most easily recognized when the view that the grounding individual duties are themselves moral is combined with the caring, we-reasoning, and preparedness accounts. Can there be moral duties to think (we-reason) in certain ways? Can there be moral duties to have certain dispositions (be prepared, care)? Traditionally, moral obligations are understood to range over actions rather than dispositions or thoughts. The traditional view fits with the fact that human persons have a kind of control – volitional control – over actions that they do not have over dispositions or thoughts. Over dispositions and thoughts, people (arguably) merely have what can be called rational control. To illustrate, you can decide to say that the earth is flat, but you cannot decide to believe that the earth is flat; rather, you form, maintain, revise, or abandon beliefs in response to (what you perceive to be) epistemic reasons. The traditional view can be strengthened by pointing out that we seem to hold each other morally responsible in more severe ways (blame, punishment) for performing morally wrong actions than we do for having bad dispositions or thoughts.[4]

The traditional view is, of course, controversial. I am not defending the traditional view. My point is merely that the accounts of collective duties under consideration, when combined with the view that the grounding individual duties are moral, is committed to rejecting the traditional view. This seems to be a high cost, at least if we are looking for an account of collective moral duties that is compatible with as many (prima facie plausible) ethical and meta-ethical positions as possible.

The we-reasoning, preparedness, and caring accounts should thus better not be combined with the view that the grounding individual duties are moral. At least not if a more attractive alternative is available. It is less clear how much of a problem this is for the collectivization account. This is because it is less clear what exactly it takes to transform an unstructured group into a group that has its own moral duties. Does it take certain dispositions or thoughts, or is the performance of certain actions sufficient? I will not try to answer this question here as this would lead us too far afield.

The next suggestion to consider is that the individual duties that ground collective moral duties are themselves not moral but (merely) rational duties. This suggestion has the advantage of being compatible with the traditional view that moral duties range over actions but not over dispositions or thoughts.

The problem with the suggestion, though, is that it comes with meta-ethical commitments regarding the relation between morality and rationality. (I here use *rationality* as synonymous with what is often called *practical reason*, as referring to what one ought (simpliciter) or has reason (simpliciter) to do, whereas morality is concerned

[4] On the traditional view and the debate surrounding it, see Portmore, 2019 and Clarke, 2023.

with what one ought morally or has moral reason to do.) Is it always rationally required to be moral? Some meta-ethical views (e.g., many versions of naturalist realism) answer this question in the negative. However, the individual duties in question ground, and are thus closely connected to, collective moral duties. Accordingly, the contents or phenomenology of the grounding individual duties suggest that these duties have a moral character. This does not sit well with meta-ethical views that detach morality from rationality.

My suggestion is a hybrid account. The individual duties that ground collective duties are rational duties of those persons of whom it is rationally required to be moral (in general or at least in the relevant situations). This hybrid account seems to have many advantages.

First, unlike the first suggestion, the hybrid account is neutral regarding the issue of whether moral duty ranges over dispositions and thoughts. If we reject the traditional view according to which moral duty ranges only over actions and instead assume that moral duty also ranges over dispositions and thoughts, we can just add that the grounding individual duties are not merely rational but also moral. On the other hand, proponents of the traditional view can agree that the grounding individual duties are rational and reject the additional claim that the duties are also moral.

Second, unlike the second suggestion, the hybrid account is neutral regarding the issue of whether it is always rationally permitted or even required to be moral. On an extreme view, we can just assume that all moral requirements are also rational requirements. Then the hybrid account simply yields that every person has the grounding individual duties.

Third, unlike the second suggestion, the hybrid account accommodates (like the first suggestion) the seemingly moral contents and phenomenology of the grounding individual duties. The hybrid account achieves this by stating that the grounding individual duties are rational duties of persons who are rationally required to be *moral*.

4. Are the Individual Duties Perspective-Dependent?

Several proponents of accounts of collective duties have committed themselves to the view that the individual duties that ground collective duties are perspective-dependent, i.e., roughly, that they depend on the individuals' epistemic situations. In this section I argue that such a commitment is mistaken.

Let us begin by looking at what Anne Schwenkenbecher (2021: 17–18) says about the issue:

> The notion of collective obligations defended here aligns best with what Zimmerman (1996) calls "the prospective view of moral obligation". […] This

> means that our moral obligations depend on our reasonable, justified (but not necessarily true) beliefs concerning the problem at hand.
>
> The prospective view of moral obligations makes better sense of the intuition that agents have no collective obligation to address a joint-necessity problem [a moral problem they can solve only together but not individually] where they reasonably believe an individually available option to be superior to an only collectively available option; or where they reasonably disagree on which collectively available option is best and they therefore cannot agree on a course of action; or where they are unlikely to figure out the collectively optimal solution in the time available to them. These kinds of complications, where they cannot easily be resolved between willing agents, can cancel collective obligations.[5]

I share the intuition highlighted by Schwenkenbecher. But should the intuition be accommodated in our theories of collective duties? I don't think so.

Intuitions (based on cases as well as other considerations) have also been put forward in areas of moral theory that do not concern collective but individual obligations (e.g., Jackson, 1991). The intuitions there are arguably not weaker than the intuition under consideration in the context of collective duties. Nonetheless, objectivists – those who claim that moral duties do not depend on the agent's epistemic situation – have defended their position against such intuitions (Graham, 2021). Such defenses do not only include attempts to accommodate or, alternatively, debunk intuitions supporting the prospective view of moral obligation, henceforth "prospectivism". Objectivist arguments also include attempts to reveal intuitions that support objectivism.

The important point is that it is currently an open question which side is correct. Accordingly, just as it would be desirable to have a theory that is neutral regarding the objectivism/prospectivism debate in the case of individual morality (say, in the ethics of promising), it would also be desirable to have a theory of collective duties that is neutral regarding the objectivism/prospectivism debate. This demand is based on the same desideratum that was employed in the previous section. Other things being equal, an account of collective duty should be compatible with as many (prima facie plausible) ethical and meta-ethical positions as possible.

This suggests that if we can formulate our accounts of collective duty in ways that are neutral regarding the objectivism/prospectivism debate, then we should choose such formulations rather than trying to adjudicate between objectivism and prospectivism. In the remainder of this section, I shall suggest that the accounts of collective duties considered here can be formulated in neutral ways.

[5] The reference in Schwenkenbecher's text is false. A fitting reference would be Zimmerman, 2008.

Let us consider again Björnsson's (2020: 132–3, footnotes and references omitted) definitions of the accounts:

> It has […] been suggested that a group's obligation to φ is grounded in the fact that the group would φ, or would be sufficiently likely to φ, if members discharged their individual obligations to:
>
> > *take steps to collectivize*: to transform the group into a group agent that has its own obligation to φ (Collins 2013; cf. Hindriks 2019; Isaacs 2011: 144–54 on "putative obligations");
> >
> > *we-reason*: to identify φ-ing as the optimal solution to a problem that group members cannot solve individually and to deduce their own individual actions based on this (Schwenkenbecher 2018; 2019);
> >
> > be prepared to do their part in φ-ing should they be sufficiently certain that others would as well (Aas 2015); or
> >
> > care to the right extent about what is morally at stake, in the sense of being disposed to (i) pick up information about what reactions and actions tend to promote what is morally important and (ii) be moved by such information when opportunity arises (Björnsson 2014, Forthcoming).

It is possible, I think, to state all this in an objectivist view. Here is a preliminary attempt. On the objectivist versions of the approaches, a group's obligation to φ is grounded in the fact that the group would φ, or be sufficiently likely to φ, if members discharged their individual obligations to

- *take steps to collectivize*: to transform the group into a group agent that has its own obligation to φ;
- *we-reason*: to (correctly) identify φ-ing as the optimal solution to a problem that group members cannot solve individually and to (correctly) deduce their own individual actions based on this;
- *be prepared* to do their part in φ-ing if others would as well; or
- *care* to the right extent about what is morally at stake, in the sense of being disposed to be moved by facts about what reactions and actions promote what is morally important when opportunity arises.

Of course, if we constantly worked with both objectivist and prospectivist versions of accounts of collective duties, this could make our texts messy. But this is a practical problem that merely concerns how we present our thoughts. This problem cannot justify stating our thoughts merely in a prospectivist framework.

5. Conclusion

I have considered three issues arising from ethical theory for accounts that ground the moral duties of unstructured groups in individual duties. The first issue concerns the moral status of the collective behavior that is meant to come out as obligatory on the accounts. Since the individual duties grounding the collective duty are specified with regard to the collective behavior, we need to clarify the moral status of the collective behavior as it features in the individual duties. After elaborating that the collective behavior better not be categorized as obligatory in the individual duties, I have argued that none of the suggestions made so far – categorizing the collective behavior in terms of moral importance, optimality, or conditional duties – is satisfactory.

The second issue concerns the nature of the grounding individual duties. After raising objections to views that understand the individual duties exclusively as moral or as rational (but not moral), I have suggested a hybrid view. According to the hybrid view, the individual duties that ground collective duties are rational duties of those persons of whom it is rationally required to be moral (in general or at least in the relevant situations).

Finally, I have argued that accounts of collective moral duties should remain neutral on the question of whether the grounding individual duties depend on the individuals' respective perspectives. I have suggested that it is possible, though perhaps cumbersome in terms of presentation, to formulate the accounts in both way. Moreover, I have indicated that there is a large and complex debate between objectivists and prospectivists and argued that it would be a mistake for theorists of collective duties to commit themselves to either position.

References

Aas, S. (2015). Distributing Collective Obligation. *Journal of Ethics and Social Philosophy*, *9*(3), 1–23. https://doi.org/10.26556/jesp.v9i3.91

Björnsson, G. (2014). Essentially Shared Obligations. *Midwest Studies in Philosophy*, *38*(1), 103–120. https://doi.org/10.1111/misp.12019

Björnsson, G. (2020). Collective Responsibility and Collective Obligations Without Collective Moral Agents. In S. Bazargan-Forward & D. Tollefsen (Eds.), *The Routledge Handbook of Collective Responsibility*. Routledge.

Björnsson, G. (Forthcoming). On Individual and Shared Obligations: In Defense of the Activist's Perspective. In M. Budolfson, T. McPherson, & D. Plunkett (Eds.), *Philosophy and Climate Change*. Oxford University Press.

Clarke, R. (2023). The Source of Responsibility. *Ethics*, *133*(2), 163–188. https://doi.org/10.1086/722124

Collins, S. (2013). Collectives' Duties and Collectivization Duties. *Australasian Journal of Philosophy*, *91*(2), 231–248. https://doi.org/10.1080/00048402.2012.717533

French, P. A. (1979). The Corporation as a Moral Person. *American Philosophical Quarterly*, *16*(3), 207–215.

Graham, P. A. (2021). *Subjective Versus Objective Moral Wrongness*. Cambridge University Press.

Hindriks, F. (2019). The Duty to Join Forces: When Individuals Lack Control. *The Monist*, *102*(2), 204–220. https://doi.org/10.1093/monist/onz006

Isaacs, T. (2011). *Moral Responsibility in Collective Contexts*. Oxford University Press.

Jackson, F. (1987). Group Morality. In J. J. C. Smart, P. Pettit, R. Sylvan, & J. Norman (Eds.), *Metaphysics and Morality: Essays in Honour of J.J.C. Smart*. Blackwell.

Jackson, F. (1991). Decision-Theoretic Consequentialism and the Nearest and Dearest Objection. *Ethics*, *101*(3), 461–482. https://doi.org/10.1086/293312

List, C., & Pettit, P. (2011). *Group Agency: The Possibility, Design, and Status of Corporate Agents* (P. Pettit, Ed.). Oxford University Press.

Portmore, D. W. (2019). *Opting for the Best: Oughts and Options*. Oxford University Press.

Portmore, D. W. (2022). Consequentializing. In E. N. Zalta & U. Nodelman (Eds.), *Stanford Encyclopedia of Philosophy* (Fall 2022 Edition). https://plato.stanford.edu/archives/fall2022/entries/consequentializing/

Rosenqvist, S. (2019). The No Act Objection: Act-Consequentialism and Coordination Games. *Thought: A Journal of Philosophy*, *8*(3), 179–189. https://doi.org/10.1002/tht3.418

Schroeder, M. (2007). Teleology, Agent-Relative Value, and 'Good'. *Ethics*, *117*(2), 265–295. https://doi.org/10.1086/511662

Schwenkenbecher, A. (2018). Making Sense of Collective Moral Obligations: A Comparison of Existing Approaches. In K. Hess, V. Igneski, & T. L. Isaacs (Eds.), *Collectivity: Ontology, Ethics, and Social Justice* (pp. 109–132). Rowman and Littlefield.

Schwenkenbecher, A. (2019). Collective Moral Obligations: 'We-Reasoning' and the Perspective of the Deliberating Agent. *The Monist*, *102*(2), 151–171. https://doi.org/10.1093/monist/onz003

Schwenkenbecher, A. (2021). *Getting Our Act Together: A Theory of Collective Moral Obligations*. Routledge.

Tännsjö, T. (2007). The Myth of Innocence: On Collective Responsibility and Collective Punishment. *Philosophical Papers*, *36*(2), 295–314. https://doi.org/10.1080/05568640709485203

Wringe, B. (2010). Global Obligations and the Agency Objection. *Ratio*, *23*(2), 217–231. https://doi.org/10.1111/j.1467-9329.2010.00462.x

Zimmerman, M. J. (1996). *The Concept of Moral Obligation*. Cambridge University Press.

Zimmerman, M. J. (2008). *Living with Uncertainty: The Moral Significance of Ignorance*. Cambridge University Press.

Olle Blomberg[1] & Björn Petersson[2]

Collective Blameworthiness and the Group's Perspective[3]

A violation of a collective moral obligation can take place without each member violating an individual obligation. That may seem problematic. A violation of a moral obligation typically justifies moral blame. If we blame a group, individual members will register the blame. According to an influential view from John Stuart Mill and others, the primary function of moral blame is to evoke feelings of guilt, and guilt feelings, as Mill says, are unpleasant and can be considered as a basic form of punishment. Also, feeling guilty involves acknowledging fault. Then, in line with the Millean view, the individual member may be punished for a violation she did not commit, and be required to take on responsibility for a fault that was not hers, which appears unfair as well as incoherent. Given the Millean view of moral blame, it seems then that we should give up the idea that groups can have irreducibly collective obligations. We confront this objection by explaining how genuine feelings of guilt which are unpleasant and involve acknowledging fault can be the appropriate response to moral blame towards one's group, even for an individually innocent group member. We thereby reconcile the Millean view of moral blame with the possibility of irreducibly collective moral obligations. Our explanation is based on the idea that an individual can identify with her group in a strong sense, and harbour guilt feelings from different perspectives.

[1] University of Gothenburg, olle.blomberg@gu.se.

[2] Lund University, bjorn.petersson@fil.lu.se.

[3] The Lund Gothenburg Responsibility Project (PI: Prof. Paul Russell), funded by the Swedish Research Council, is gratefully acknowledged.

1. Introduction

We often target several individuals considered collectively, as a group, with blame for outcomes they fail to prevent as well as for collective wrongdoing. The group might consist of a few individuals, as when we blame a group of co-workers for the bullying of a colleague. Or the group might consist of a large mass of people, as when we blame a state's citizens for electing an authoritarian leader or when we blame the world's affluent people for failing to slow down global warming. We have previously argued that at least in the former sort of small-scale case, several agents can have an irreducible collective moral obligation (Blomberg & Petersson 2023).

In this paper, we discuss what happens when such a group fails to act in accordance with their obligation: what sort of blaming responses may aptly be directed to individual members of such a group? We are particularly interested in cases where one or more members may not individually be at fault even though the group is blameworthy for violating an obligation.

Drawing inspiration from work in social psychology on group identification and building on our account of collective moral obligation, we argue that it can be morally fitting for group members, including those not individually at fault, to feel guilt from the group's perspective, in light of the group's failure to act in accordance with its obligation. Moreover, one function of directing moral blame towards a collective may be to evoke such we-feelings of guilt in the individual members. Our view that it can be fitting for members to feel guilt from the group's perspective gives substance to the idea that an unstructured group, what Virginia Held (1970) calls "a random collection of individuals", can be collectively blameworthy in a way not reducible to some aggregation of the individual blameworthiness of members. It is not uncommon to assume that a socially organized group may be a proper target of moral blame, on the condition that it has an established collective decision procedure or fulfils some other substantial criterion meant to capture what is necessary for a group to possess moral agency in its own right. Companies and other organizations are paradigmatic examples. Like Held, we instead focus the conditions under which a group lacking such organizational features can be blameworthy in a non-distributive sense.

We compare and contrast our view to accounts claiming that assignments of collective guilt have no implications at all for individual members' guilt (Cooper 2001; Gilbert 2002), to views according to which collective guilt is fitting because guilt does not imply fault (Morris 1988; Sepinwall 2011; Velichkov 2023), and to views according to which some other moral emotion than guilt is fitting in those members not individually at fault (Oshana 2006; Telech 2022)[4]. While we have considerable

[4] Gunnar Björnsson seems to hold a view of the last kind as well: A group can be retrospectively morally responsible for a bad outcome without each group member having a substandard quality of will (2014: 113-114,

sympathy for views of the last kind, we here seek an account of collective blameworthiness that provides a direct response to the common individualistic objection that a practice that takes collective responsibility seriously will necessarily be unfair and/or incoherent.

2. Collective moral obligation

When agents only together can bring about an outcome that is morally good or best, and they know that they can do this without significant risk to themselves, they will at least sometimes be morally required to bring about the outcome (Blomberg & Petersson 2023; McKinsey 1981; Copp 1991; Wringe 2010; Cripps 2013; Björnsson 2014; Pinkert 2014; Aas 2015; Schwenkenbecher 2021). In such cases, it will be to *them*—collectively rather than distributively—that a moral demand to bring about the outcome should be directed.

On our account, several moral agents have such a moral obligation together only if they each have (i) a context-specific capacity to view their situation from the group's perspective—to "group identify"—and (ii) at least a general capacity to deliberate about what they ought to do together (Blomberg & Petersson 2023; see also Schwenkenbecher 2021). If they also have the joint ability to realize a morally required outcome, then they together have the collective obligation to bring it about.[5] An agent who identifies with her group in our sense does not merely see herself as a member of the group and does not merely care about the group.[6] A group-identifying member also views the choice situation from the group's perspective. If she also ends up deliberating about what they should do—engaging in so-called "team reasoning" (e.g. Bacharach 2006; Colman & Gold 2018)—then she evaluates different courses of action open to the group and infers that she, as a group member, ought to do her part of the optimal action profile available to the group. If an agent lacks capacities for group identification and team reasoning, then she will not be able to grasp the normative reasons that make the collective action morally required. Compare: a singular individual would not have a moral obligation to raise and take care of a child unless she had capacities to identify as a persisting person over time and for planning and coordinating her activities over time. Without these capacities, she would not be able

fn. 10), and it would not be fitting for those with a satisfactory quality of will to feel guilt (2021: 3556). Instead, for them, "sadness, pain, horror, disappointment, contempt, or shame" may be fitting (ibid.).

[5] For details, see (Blomberg & Petersson 2023). For critical discussion of our account, see (Ludwig 2023).

[6] The notion of group identification is central in social identity theory (Hogg et al. 2017). Some empirical findings suggest that group identification prompts cooperation and team reasoning in social dilemmas and coordination games, although the data is hardly conclusive or decisive (see Thom et al. 2022).

to grasp the normative reasons that make the cross-temporal activity of raising and taking care of the child morally required for her.

We distinguish between an individual obligation, understood in terms of what I ought to do given my expectations about the actual behaviour of people around me, and what we might call a participatory obligation, which is what I ought to do as part of what we ought to do.[7] Wilfrid Sellars notes that even when my individual action is what figures in the content of my intention, the intention can be held either from my group's perspective or from my individual perspective. "[We] can say that Jones intends to do A *sub specie* 'one of us,' and flag our representation of his intention with a subscript 'we,' thus, Jones intends 'Shall$_{we}$ [I do A]'" (Sellars 1980: 99). To paraphrase Sellars, we can say that I have a participatory obligation to do A when I ought *sub specie* "one of us" to do A.

An individual agent's participatory obligation and her individual obligation may be in irresolvable practical conflict. Consider a moral multi-player social dilemma such as the following:

> *Community School*: In our small local community, I can either pay a school tax or keep my money to pay for some private teaching for my children. Each parent in our community has this choice. My children would get excellent education if they could go to our public school and get extra private teaching on top of that, good education if they merely go to our community school, considerably less good education if they merely get the private teaching that I can afford, and no education at all if I can't pay for private teaching and there is no public school. A sufficient number of school tax payers are required to sustain the school. The number of school tax payers in our community is far below that threshold. Therefore, I would merely make the situation worse for my children by paying school taxes, and paying them would not make the situation better for anyone else. The same is true of each parent in our community.

Arguably, I would not violate any individual obligation to my children or to anyone else by not paying the school tax. My children would be worse off if I made myself unable to pay for private teaching. But we parents together give our children a considerably less good education than we could have given them. Given that we each have a context-specific capacity to view our situation from the perspective of the community of parents, as well as at least a general capacity to deliberate about what we ought to do together, then we would arguably together violate a collective obligation

[7] This distinction between an agent's individual obligation and his or her participatory obligation is not explicitly drawn in (Blomberg & Petersson 2023).

to sustain the public school. In this kind of community, conditions for group identification may be more or less favourable. The plausibility of holding the group collectively to account for violating an obligation will vary with those conditions.

We do not think that this type of dilemma is uncommon in real life, nor that the feeling of being torn between the duties they typically incur is irrational. In such cases, there is an irreconcilable conflict between an individual obligation—what the individual agent ought, from the individual perspective, to do, and a participatory obligation—what the individual agent ought, from the perspective of the group with which the agent identifies, to do. (One might think that each parent would violate an individual obligation to engage and convince the others to together pay enough school taxes to sustain the school by e.g. handing out pamphlets, talking to friends and neighbours etc. It is possible that each knows that this would be fruitless though, in which case no one would have such a complex individual obligation. And even if each has such a complex individual obligation, these individual obligations can co-exist with the collective moral obligation of the group and the ensuing participatory obligations.)

On our account, a collective moral obligation is not reducible to individual moral obligations. Individual agents can together, as group members, violate their collective obligation without all or any of them violating any individual obligations. Nevertheless, a collective obligation does not require that the group in question has a formal organization or that it is a unified group agent in some substantial sense. What is required is rather that the group members in the specific decision context can identify with the group and can frame their options from a joint plural perspective.

3. Blame and the individualist's objection

In cases where the subject of an obligation violates it and lacks excuse, it is appropriate for the victims of the violation, or for members of the moral community at large, to direct blame towards the perpetrator. According to some, the core function of blame is to protest wrongdoing (e.g. Hieronymi 2001; Talbert 2012). According to others, it is to signal one's endorsement of the moral norm that has been violated (Shoemaker & Vargas 2021). Yet others take the core function of blame to be that of getting the target to blame themselves and to evoke feelings of guilt in the target (e.g. Mill 1863: 33; Brandt 1958: 16–17; Gibbard 1990: 150). In this paper, we do not take a position on what the defining core function of blame is, nor do we take a position on whether it at all has such a core function. Instead, we assume for the sake of argument at least that in many cases where we express and communicate blame, we do that with the aim of evoking guilt feelings in the target. We make this assumption because it forms

the basis of a common objection to the very idea of collective blameworthiness that we want to respond head on to.

A group member can be 'individually innocent' to her group's wrongdoing in the standard sense that she did not intentionally make a marginal contribution to it. Given that the individual group members are the ones who will hear and register the moral complaints against the group that has violated the collective obligation, evoking guilt feelings in the group may seem unfair at least to those members who have not violated any individual obligations. If the individually innocent member acknowledges the moral blame directed towards his group, he will be punished "by the reproaches of his own conscience" (Mill 1987 [1863]: 65). Hence, to continue using John Stuart Mill's language, the member will be punished for a violation he has not committed.

Moreover, as Galen Strawson (1994: 9) points out, it is doubtful whether it is conceptually possible to feel guilt for something without believing that one is morally responsible for it. So, if some members of a group have no reason to believe that they are guilty of any individual wrongdoing, directing blame towards the group may not only seem unfair, but nonsensical since it aims at evoking incoherent attitudes in its recipients (cf. Wallace 1994: 135).

A group can fail to fulfil a collective moral obligation even if some or all of its members have fulfilled their individual obligations, or so we and others have claimed. In light of the two potential problems with collective blame – unfairness and incoherence – what would be the appropriate moral attitude towards such a group? And how ought an individually faultless group member respond to blame directed at her group?

Some, like Margaret Gilbert, have argued that assignments of collective guilt have no implications for individual members' guilt (Gilbert 2002; Cooper 2001). But given Mill's and others' observation that an important function of moral blame is to evoke feelings of guilt, and that groups as such arguably are incapable of harbouring such feelings, it is difficult to see the point of blaming a group as such without implying anything about its members. Moreover, even if there was some plausible sense in which groups as such *were* capable of feeling guilt, as Gilbert claims, it is difficult to see how moral blame towards the group could avoid affecting group members in addition to that, and make them react with the unpleasant feelings that blame normally creates. This needs to be justified.

Others have claimed that appropriate guilt reactions do not imply fault (e.g. Morris 1988; Sepinwall 2011; Velichkov 2023: 59–80). One may e. g. feel guilty for merely having certain bad thoughts or being a survivor of some catastrophe ("survivor guilt") (Morris 1988). This move makes it easy to accommodate that group members who are not themselves at fault may nevertheless appropriately feel guilty for what

others in their group have done. Admittedly, these kinds of painful reactions – feeling guilty for bad thoughts, feeling survivor guilt, and feeling guilty, e.g., for what a family member has done – seem psychologically natural, understandable, and they are probably not unusual. However, one may question whether such guilt feelings are really *appropriate* unless there is *some* reason to think that the thoughts, the survival, or the family member's behaviour, is connected to a moral shortcoming of the agent, like being disposed to act on the bad thoughts, failing to help others in need to save oneself, or failing to intervene to set one's family member straight. In cases where we exclude the possibility of such connections, we typically try to talk people out of having such painful feelings, or recommend that they get professional help to get rid of them, and intuitively this seems to be the right thing to do.

Thirdly, in the context of collective blameworthiness some have rejected the view that guilt feelings must be the appropriate responses to moral blame directed at the group (e.g. Oshana 2006; Björnsson 2021; Telech 2022; Knudsen 2023). Blame, on this view, can call for a variety of reactive attitudes. Members who are not at fault may appropriately feel, say, shame, a kind of agent-regret, or disappointment. This seems plausible and there is no reason to deny that blame may have multiple functions.

However, our aim here is to explain how moral blame towards a group can make sense *even* on the Millean assumption that blame aims at guilt feelings and thereby constitutes a basic form of punishment.[8] That assumption is the basis for aversiveness to the idea of collective blame and responsibility, and it is what gives rise to worries about unfairness. Moreover, we should at least admit with Allan Gibbard and others that blame and guilt as characterised by Mill occupy a central "region in our moral thought" (Gibbard 1990: 52). Then it seems important to examine to what extent that strong conception of moral blame is applicable in the collective context. Finally, the move to less paradigmatic conceptions of blame appears *ad hoc* when we analyse what people do when they blame groups, insofar as the move is motivated solely by a concern that collective guilt and punishment seems unpalatable to many. Therefore, our ambition is to provide an account of moral blame towards collectives that retains the Millean assumptions about the functions of moral blame and guilt feelings. Such an account should explain, rather than reject, the connection between (collective) guilt and (collective) fault, as well as the connection between acknowledging collective wrongdoing and feeling guilt as a member.

[8] For attempts to make sense of moral blame directed at collectives without the Millean assumption, see (Garcia 2022; Smith 2009).

4. Collective guilt

When a collective obligation is violated, several agents are jointly blameworthy.[9] How does our account then deal with the problems associated with collective blame mentioned earlier (unfairness and incoherence)? While the solution we give here fits our account of collective obligation especially well, it can in principle be combined with other accounts of collective obligation.

On our account, moral obligations are always relative to an agential ("I"/"we") perspective. Similarly, blameworthiness is always relative to such an agential perspective. The behaviour of several agents can be framed as a violation of a collective obligation, or it can be viewed as the combined result of several agents each reasoning and acting relative to what they each, considered as individuals, ought to have done. When we blame the group as a collective, we address its members in a way that is different from what we do when we blame them as individuals. We address them as parts of the group to which we ascribe the violation of the collective obligation. This holds true also when the one blaming is herself a member of the group, engaging in plural self-blame: If she identifies with her group, she may feel guilty from its perspective - she may have "we-feelings" of guilt (Petersson 2020). The motivational role of the feeling evoked by registering blame towards one's collective may differ from the role of the feeling which is a response to being blamed as an individual.[10]

We typically direct blame towards groups as such when we want to stress the collective character of the action or omission for which the group is blamed. Our suggestion is *not* just that a member can feel guilt *for* her group or what it has done, in analogy with how one may feel embarrassed or ashamed for someone else's behaviour. An individual who merely categorizes herself as a group member but does not identify with the group in our strong sense can feel such vicarious guilt for what

[9] Rowan Mellor (2024) argues that while there are collective obligations, only an individual is a fitting target of resentment. Given Millean assumptions about blame (which Mellor may reject), this would imply that while several agents can jointly violate an obligation, individuals would at most be severally blameworthy. An individual is blameworthy in the sense of being the fitting target of a victim's resentment on Mellor's view only if they act "out of a lack of due concern for their legitimate interests." (61) However, the object of blameworthiness and resentment cannot in that case be the violation of the collective obligation, but only the individual's violation of her own obligation not to act out of a lack of due concern or to do her part (or some more complex conditional individual obligation; see e.g. (Goodin 2012; Collins 2019: 116–117)). We therefore think that Mellor's view risks undermining the claim that there are collective obligations.

[10] Our account has affinities with Nicolai Knudsen's (2023) account of appropriate collective blame. Knudsen does not single out guilt as the proper response from members to such blame but he mentions that we expect a member, M, who is individually innocent "to see and measure himself in light of the failed group effort and, as a result, to direct negative reactive attitudes toward himself, e.g., to feel guilty, regretful, or ashamed about the group failure." (164). Furthermore: "If we believe that M's relation to the failed group effort warrants such self-directed negative reactive attitudes, we blame M as a group member." (ibid.) And we can blame M, and M can blame himself, both "*tout court*"—as an individual—and as a group member (150).

the group has done. Feeling guilty for what we have done *from our perspective* involves group identification in our strong sense though. There is a motivational difference between these ways of feeling guilt in light of the wrongdoing of one's group.[11] As we have argued elsewhere, this motivational difference can be brought out by considering certain problematic social choice situations, as Michael Bacharach's (2006) work on team reasoning indicates. To simplify: Guilt feelings from our perspective do not only trigger thoughts about what I ought to have done for us, or what I ought to have done given my expectations about what others would do, but primarily thoughts about what *we* ought to have done.

To summarize, when we assign an irreducibly collective obligation to a group, we implicitly assign a capacity for group identification to its members (Blomberg & Petersson 2023). Our account of collective moral blame mirrors this claim about collective obligations. Such blame appeals to a presumed capacity of group members to feel guilt from the group's perspective. In this way, we don't have to give up any of the standardly assumed connections between collective guilt and membership guilt, between feeling guilty and acknowledging wrong-doing, or between moral blame and genuine guilt feelings.

We-feelings of guilt are just as genuine as I-feelings of guilt on this view. Genuine guilt feelings are painful or at least unpleasant, and this is part of what explains the motivational importance of guilt feelings. So, will not the individually innocent member who feels guilty from her group's perspective for what her group has done or failed to do be unfairly punished, as Mill says, by the reproaches of her own conscience, just as much as someone who feels guilty for something she is no part of at all?

But recall, firstly, that on our account blame towards a group is fitting only given the assumption that the members of the group had at least a capacity to identify with the group in the situation where the group failed. A person's capacity to identify with a group of people may depend on various external factors as well as on social cues which in turn may have some moral relevance.[12] Moreover, such factors are relevant

[11] Empirical research in social psychology and the social sciences on "collective guilt" (Goto & Karasawa 2011; Ferguson & Branscombe 2014) or "group-based guilt" (Hakim et al. 2021) does not always distinguish between vicarious guilt for what a group that one is associated with has done, and feeling guilt for what a group has done from the group's point of view—that is, from the retrospective deliberative point of view that concerns what the group ought to have done. But some of this empirical research may be measuring the phenomenon we are interested in: we-feelings of guilt.

[12] It would reach too far to attempt to provide a proper account of the concept of 'capacity' here. We want to stress two things though. First, we are not only referring to the general capacity to identify with a group, presumably shared by all normally functioning adult human beings, but to an ability to identify with a group when in a relatively specific concrete choice context (a useful discussion about how to think about more or less specific abilities/capacities and how they relate to relevant contexts can be found in (Jaster 2020, especially section 4.5)). The question of whether someone has the capacity to group identify in a certain situation plausibly depends on

to how the blameworthy group should be delimited and separated from innocent bystanders to begin with.

Consider the following situation:

> *Workplace:* Agnetha, Björn and Benny are good friends and work closely together in a research team. Frida is employed within the same team and performs her work well but rarely talks to the others or takes part in their social activities. There is a growing shared sense between Agnetha, Björn and Benny that Frida is different from them, and a bit odd. Björn and Benny even start to make fun of Frida at times, while Agnetha is careful to treat Frida respectfully at all times. At some point, Frida cannot avoid being affected by the situation. This makes Frida sad and eventually severely depressed. When Agnetha, Björn and Benny realise how the situation has affected Frida, they all feel guilty about it.

Suppose Agnetha did what she had good reasons to think would be least hurtful to Frida throughout this whole process. Is it rational of her to feel guilty? Would it be correct to blame Agnetha for Frida's getting depressed? This is a case where it seems likely that Agnetha identifies with the group of three and that there is "a sense of us" – that is an element in the group's sense of there being something odd about Frida. Moreover, Björn's and Benny's making fun of Frida is partly a manifestation of an attitude towards Frida that Agnetha shares with Björn and Benny, an attitude that none of them might have acquired as individuals without the dynamics within this specific group.

In line with our previous account, we find it natural for Agnetha to feel guilty as a group member, to feel that *we* have wronged Frida – i.e. to feel guilt from her group's perspective. Moreover, it would seem at least permissible to criticise Agnetha morally if she did not in any way react like this when she realised how Frida had been affected. This is acceptable, we think, because in this case it is obvious that all conditions for group identification are fulfilled, and also because the very formation of the group, with the diverging attitudes towards outgroup members which are typical for strong group identification, is morally relevant. At the same time, it is clear that Agnetha's individual fault – understood as her individual intentional marginal contribution to Frida's predicament – is much less grave than Björn's or Benny's, perhaps even negligible.

A natural objection to this might be to say that it may be proper of Agnetha to feel guilty for not having done enough on her own to stop the bullying, but that she

interpersonal factors ranging from verbal communication to more subtle social signs, as well as on the nature of the decision problem and on various other circumstances, some of which may be wholly external to the group. Second, the question of whether the conditions in a specific case are such that individuals are capable of group identification is empirical, and a matter of degree rather than all-or-nothing.

should not feel guilty at all if she did everything she could. We think that the dichotomy presupposed in this kind of response fails to recognise our capacity to frame situations from different perspectives. Agnetha may know that she did everything she could for Frida, given how her fellow group members acted, and she might even be justified in thinking that she cannot be blamed on account of her individual contribution to what happened. Still, it may not be improper of Agnetha to apologise to Frida on behalf of the group, and, in light of what happened, to think of future interactions in terms of how *we* should behave to make up for what we did, as opposed to thinking merely of how *I* should behave given the behaviour of others. This may also result in different attitudes and behaviour toward Björn and Benny. Instead of viewing their bad behaviour from the point of view of a bystander, and feeling indignation toward them as a third party, she will see them as co-participants who must be brought to behave in line with how they together ought to behave: she may try to get them to feel guilty from their group's point of view too, and urge them to apologise on the group's behalf as well. In other words, by framing the past situation from her group's perspective and feeling guilty from the group's point of view, she will be motivated in a way distinct from how she would be if she just considered her individual contributions.

Here is another case where it might be clearer that an individual who can fittingly feel we-guilt has done everything he could reasonably be expected to do to avoid becoming individually blameworthy:

> *Hooligans*: Patrick is a devout supporter of Brumlington Football Club. He and other Brumlington F.C. fans are travelling in the same train to Harchester to see their team play Harchester United. During the train journey, the increasingly exuberant and heedless group of supporters start to vandalise the interior of Patrick's train car. Patrick tells some of them to stop it, and even tries to physically restrain one of them, but to no avail. The other supporters either ignore him or hold him back. They end up completely demolishing the car's interior before the train arrives at Harchester station. While Patrick is convinced that he individually did all that could reasonably be expected of him in the train car, he nevertheless feels guilty for what he and the other supporters did to the train car's interior.[13]

[13] Various people have suggested to us that in this kind of case, where it is obvious that the other group members won't do their parts, each individual (including someone like Patrick) must lack the appropriately context-specific capacities to group identify and team reason about what they ought to do (for another case of this kind, see Blomberg and Petersson 2023: 23-24). It would follow that the group could not have and violate a collective obligation in this sort of case. After all, the relevant team-reasoning capacity must be a capacity for *valid* practical reasoning, and one might think that the known non-compliance of others would make team reasoning invalid (thanks to Niels de Haan for pressing this objection). But we do not find this obvious. A person's capacity to identify with a group in a specific situation depends on various cues (from things like shared history and common

Despite his individual innocence, given his social identity as a Brumlington F.C. fan, Patrick might nevertheless identify as a member of the same group as the other fans in the train car who together vandalise the train car. It might therefore be fitting for him to feel guilt from the perspective of the group consisting of himself and the other supporters in the train car, even if he is not individually morally responsible or blameworthy for contributing in any way to the vandalism. In moments when he feels estranged from other Brumlington F.C. fans, he might fittingly feel no guilt at all for what the others did in the train car, especially in light of his courageous attempt to stop the hooligans in the train car.

Like Bacharach (2006), we find it very reasonable to think that we sometimes vacillate between framing situations from our group's perspective and framing them from our individual perspective. In *Workplace*, there would be no inconsistency in Agnetha vacillating between on the one hand feeling guilty from the group's perspective and on the other hand taking comfort in knowing that her individual marginal contribution to the harm done was negligible, and feeling innocent from that perspective. We don't think that this kind of predicament is very unusual, nor that it is morally problematic. There is nothing unfair about Agnetha feeling bad about what *we* have done to Frida, and Agnetha's we-feelings of guilt are fully consistent with her feeling innocent in terms of her individual behaviour. Similarly, if Patrick in *Hooligans* group-identifies with the other Brumlington F.C. fans in the train car, then there is nothing unfair about him having we-feelings of guilt for what *we*—I and the other Brumlington F.C. fans—have done to the train car.

If we, as outsiders, view these individually innocent agents as parts of the group that has violated the collective obligation—the group of colleagues or the hooligans in the train car—then we will expect them to feel guilty from their group's point of view, to apologise or to try their best to get their fellow group members to acknowledge the collective fault and also have we-feelings of guilt. As Knudsen points out, an individually innocent group member who "was unaffected by the group's failure… and merely shrugged it off saying that he did everything he could" could rightly be viewed with "suspicion and resentment" (Knudsen 2023: 164). Of course, when an individual like Agnetha or Patrick reacts against her/his group's collective wrongdoing, this will sometimes make us regard that individual as an innocent member with respect to the group's particular wrongdoing—that it, not regard him or her as

interests to more subtle inputs like body language, proximity and physical environment), not only on whether the others are likely to do their parts in what the person identifies as the best collective alternative. And we see no reason to think that a team member's capacity to validly team reason—i.e. to reason in a valid way about what is the best option available to the team together and what his own part of it is—must be obstructed by knowledge that the behaviour of the other members will result in a collective failure to realise the best option. From the point of view of the team-reasoning member, the mistake would be in the collective action rather than in his or her reasoning.

part the group to which we ascribe the collective failure. The individual may come to regard herself as such an innocent member too, for the same reason. She may then still categorise herself as a member and care about the group's moral worth. She may feel disappointment, regret or shame for the group's collective wrongdoing, and perhaps even feel vicarious guilt for it. However, she will not have a we-feeling of guilt for it unless she identifies with the group in our strong sense.

In cases where a group member reacts against her group's collective wrongdoing, her capacity for continuing to identify with the group may erode. Since group identification may be prompted by a variety of different factors apart from the members' attitudes to a specific decision, this consequence does not follow by default though. Hence, she may retain the context-specific capacity to group identify and therefore be a potential addressee of fitting collective blame (from herself or others) that targets her qua member with potential for group identification, even if she is in fact seen (by herself or others) as an innocent member who does not identify with her group in our strong sense.

5. Conclusion

The general point we have made previously is that when we direct a moral demand to a group as such by assigning an obligation to it we appeal to a capacity of each to regard their situation from a group perspective and to deliberate about what they ought to do—to team reason (Blomberg & Petersson 2023). Each having such a capacity is a necessary condition for the agents to have a collective moral obligation to begin with. We now suggest that we can make sense of blame directed at a group for failing to fulfil its obligation as involving an appeal to this same capacity. Moral blame directed towards a group when it has failed to fulfil an obligation can appeal to the members' capacities to view their failure from the collective perspective and have we-feelings of guilt. This account of collective blame, we have argued, can make sense of the possibility of fitting collective blame even given the Millean assumption that blame aims at guilt feelings and thereby constitutes a basic form of punishment.

We have argued that the relevant capacity for group identification is tied to the relatively specific context of choice in which the group violated its obligation, and that the external conditions that the individual agents were in must have been relatively favourable for group identification. Favourable external conditions could have been that the group were confined to a limited common space, that there were no obstacles to communication or signalling preparedness to coordinate, and so on.

What about large-scale collectives, like those of the world's population who contribute to global warming? The question of whether it makes sense to blame such a group morally, or regard it as proper for members of such a group to feel guilty from

their group's perspective, will depend on empirical matters concerning the actual conditions for group identification in which the group failed to live up to presumed collective obligation. If those conditions made it unlikely for a member of such a group to ever be able to identify with the group, then our view implies that it makes little sense to demand collective guilt feelings, i.e. individual we-feelings of guilt. In that case, such large-scale problems may be more effectively handled in terms of political responsibility than in terms of moral guilt.

More favourable external conditions for group identification makes moral blame for violation of an obligation to act collectively more apt. In considering thought experiments as well as real life cases, it seems to us that the strength of our intuitions about several agents' collective blameworthiness vary with the presence or absence of features favouring group identification, in a way predicted by our theory.[14]

References

Aas, S. (2015). Distributing Collective Obligation. *Journal of Ethics and Social Philosophy 9*(3), 1–23. https://doi.org/10.26556/jesp.v9i3.91

Bacharach, M. (2006). *Beyond Individual Choice: Teams and Frames in Game Theory* (N. Gold & R. Sugden, Eds.). Princeton University Press.

Björnsson, G. (2014). Essentially Shared Obligations. *Midwest Studies in Philosophy 38*(1), 103–120.

Björnsson, G. (2021). Being implicated: On the fittingness of guilt and indignation over outcomes. *Philosophical Studies 178*(11), 3543–3560. https://doi.org/10.1007/s11098-021-01613-4

Blomberg, O., & Petersson, B. (2023). Team Reasoning and Collective Moral Obligation in advance. *Social Theory and Practice*. https://doi.org/10.5840/soctheorpract2023120177

Brandt, R. (1958). Blameworthiness and Obligation. In Melden (Ed.) *Essays in Moral Philosophy.* University of Washington Press.

[14] We are grateful for challenging and insightful comments and questions from audiences at the Practical Philosophy and Political Theory seminar at Gothenburg University (Nov. 16, 2022), the Social Ontology 2023 conference in Stockholm, the ESPP 2023 conference in Prague, the Coordination Ethics workshop at the Institute for Future Studies in Stockholm (Sept. 27-28, 2023), a Center for Subjectivity Research seminar in Copenhagen (Oct. 31, 2023), and at the online *1o Encontro de Ontologia Social e Intencionalidade Coletiva* (April 12, 2024). Thanks also to Tim Campbell for helpful written comments on an earlier draft.

Collins, S. (2019). *Group Duties: Their Existence and Their Implications for Individuals*. Oxford University Press.

Colman, A. M., & Gold, N. (2018). Team reasoning: Solving the puzzle of coordination. *Psychonomic Bulletin & Review 25*(5), 1770–1783. https://doi.org/10.3758/s13423-017-1399-0

Cooper, D. (2001). Collective Responsibility, 'Moral Luck,' and Reconciliation. In A. Jokić (Ed.), *War Crimes and Collective Wrongdoing: A Reader* (pp. 205–215). Blackwell.

Copp, D. (1991). Responsibility for Collective Inaction. *Journal of Social Philosophy 22*(2), 71–80. https://doi.org/10.1111/j.1467-9833.1991.tb00039.x

Cripps, E. (2013). *Climate Change and the Moral Agent: Individual Duties in an Interdependent World*. Oxford University Press.

Ferguson, M. A., & Branscombe, N. R. (2014). The social psychology of collective guilt. In C. von Scheve & M. Salmela (Eds.), *Collective Emotions: Perspectives from Psychology, Philosophy, and Sociology* (pp. 251–265). Oxford University Press.

Garcia, A. G. (2022). The Unfairness Objection to the Practice of Collective Moral Responsibility. *The Journal of Value Inquiry 56*(4), 627–642. https://doi.org/10.1007/s10790-021-09795-0

Gibbard, A. (1990). *Wise Choices, Apt Feelings.* Oxford University Press

Gilbert, M. (2002). Collective Guilt and Collective Guilt Feelings. *The Journal of Ethics*, *6*(2), 115–143. https://doi.org/10.1023/A:1015819615983

Goodin, R. E. (2012). Excused by the unwillingness of others? *Analysis*, *72*(1), 18–24. https://doi.org/10.1093/analys/anr128

Goto, N., & Karasawa, M. (2011). Identification with a wrongful subgroup and the feeling of collective guilt: Subgroup identification and collective guilt. *Asian Journal of Social Psychology 14*(4), 225–235. https://doi.org/10.1111/j.1467-839X.2011.01348.x

Hakim, N., Branscombe, N., & Schoemann, A. (2021). Group-Based Emotions and Support for Reparations: A Meta-analysis. *Affective Science 2*(4), 363–378. https://doi.org/10.1007/s42761-021-00055-9

Held, V. (1970). Can a Random Collection of Individuals be Responsible? *Journal of Philosophy 67*(14), 471–481.

Hieronymi, P. (2001). Articulating an Uncompromising Forgiveness. *Philosophy and Phenomenological Research*, *62*(3), 529–555. https://doi.org/10.1111/j.1933-1592.2001.tb00073.x

Hogg, M. A., Abrams, D., & Brewer, M. B. (2017). Social identity: The role of self in group processes and intergroup relations. *Group Processes & Integroup Relations 20*(5), 570–581. https://doi.org/10.1177/1368430217690909

Jaster, R. (2020). *Agents' Abilities*. De Gruyter. https://doi.org/10.1515/9783110650464

Knudsen, N. K. (2023). A Pluralist Approach to Joint Responsibility. *Philosophy & Public Affairs 51*(2), 140–165. https://doi.org/10.1111/papa.12232

Ludwig, K. (2023). Collective Obligations and the Moral Hi-Lo Game. In A. Garcia, M. Gunnemyr & J. Werkmäster (Eds.), *Value, Morality & Social Reality: Essays dedicated to Dan Egonsson, Björn Petersson & Toni Rønnow-Rasmussen* (pp. 251-274). Lund: Department of Philosophy, Lund University.

McKinsey, M. (1981). Obligations to the Starving. *Noûs*, *15*(3), 309–323. https://doi.org/10.2307/2215435

Mellor, R. (2024). Joint Ought. *Philosophy & Public Affairs*, *52*(1), 42–68. https://doi.org/10.1111/papa.12252

Mill, J. S. (1987 [1863]). *Utilitarianism*. Prometheus Books.

Morris, H. (1988). Nonmoral guilt. In F. Schoeman (Ed.), *Responsibility, Character, and the Emotions: New Essays in Moral Psychology* (pp. 220–240). Cambridge University Press.

Oshana, M. A. L. (2006). Moral taint. *Metaphilosophy*, *37*(3–4), 353–375. https://doi.org/10.1111/j.1467-9973.2006.00437.x

Petersson, B. (2020). Collective Guilt Feelings. In S. Bazargan-Forward & D. Tollefsen (Eds.), *The Routledge Handbook of Collective Responsibility* (pp. 228–242). Routledge.

Pinkert, F. (2014). What We Together Can (Be Required to) Do. *Midwest Studies in Philosophy*, *38*(1), 187–202.

Sellars, W. (1980). On Reasoning About Values. *American Philosophical Quarterly* 17(2), 81–101.

Schwenkenbecher, A. (2021). *Getting Our Act Together: A Theory of Collective Moral Obligations*. Routledge.

Sepinwall, A. (2011). Citizen Responsibility and the Reactive Attitudes: Blaming Americans for War Crimes in Iraq. In T. Isaacs & R. Vernon (Eds.), *Accountability for Collective Wrongdoing* (1st ed., pp. 231–260). Cambridge University Press. https://doi.org/10.1017/CBO9780511976780.010

Shoemaker, D., & Vargas, M. (2021). Moral torch fishing: A signaling theory of blame. *Noûs 55*(3), 581–602. https://doi.org/10.1111/nous.12316

Smith, T. H. (2009). Non-Distributive Blameworthiness. *Proceedings of the Aristotelian Society 109*(1), 31–60. https://doi.org/10.1111/j.1467-9264.2009.00257.x

Strawson, G. (1994). The Impossibility of Moral Responsibility. *Philosophical Studies 75*(1/2), 5–24.

Talbert, M. (2012). Moral Competence, Moral Blame, and Protest. *The Journal of Ethics 16*(1), 89–109. https://doi.org/10.1007/s10892-011-9112-4

Telech, D. (2022). Relation-Regret and Associative Luck: On Rationally Regretting What Another Has Done. In A. Szigeti & M. Talbert (Eds.), *Morality and Agency: Themes from Bernard Williams* (pp. 233–264). Oxford University Press.

Thom, J. M., Afzal, U., & Gold, N. (2022). Testing team reasoning: Group identification is related to coordination in pure coordination games. *Judgment and Decision Making 17*(2), 284–314. https://doi.org/10.1017/S1930297500009116

Velichkov, A. (2023). *Responsibility and Ambivalence*. Lund University. https://portal.research.lu.se/files/158082745/Velichkov_Responsibility_and_Ambivalence.pdf

Wallace, R. J. (1994). *Responsbility and the Moral Sentiments*. Harvard University Press.

Wringe, B. (2010). Global Obligations and the Agency Objection. *Ratio 23*(2), 217–231.

Mark Budolfson[1]

Why Morality and Other Forms of Normativity are Sometimes Dramatically Directly Collectively Self-Defeating[2]

In a prisoner's dilemma, if everyone follows the strategy of self-interest, then everyone is certain to be worse off from the perspective of self-interest than if everyone had not followed self-interest instead. This shows that self-interest is sometimes directly collectively self-defeating, because it shows that sometimes everyone has all the relevant information, correctly follows self-interest, but thereby ends up worse off from the perspective of self-interest than they would have been if they had all followed some other antecedently identifiable strategy instead. In Reasons and Persons and On What Matters, Derek Parfit argues that it is a constraint on any plausible moral theory that morality must never be directly collectively self-defeating, and he claims that the most plausible versions of consequentialism, contractualism, and Kantian ethics all imply that morality is never directly collectively self-defeating. Some theorists not only agree with Parfit that morality can never be directly collectively self-defeating, but also believe that rationality and other forms of normativity can never have that property either. I argue against these theorists, with examples that show that morality and all other interesting forms of normativity are sometimes directly collectively self-defeating.

[1] Department of Philosophy, Department of Geography and the Environment, Population Wellbeing Initiative, University of Texas at Austin, mark.budolfson@austin.utexas.edu.

[2] Funding from Riksbankens Jubileumsfond in support of grant number: P22-0662 is gratefully acknowledged.

In a prisoner's dilemma, if everyone follows the strategy of self-interest, then everyone is certain to be worse off from the perspective of self-interest than if everyone had not followed self-interest instead. This shows that self-interest is sometimes *directly collectively self-defeating*, because it shows that sometimes everyone has all the relevant information, correctly follows self-interest, but thereby ends up worse off from the perspective of self-interest than they would have been if they had all followed some other antecedently identifiable strategy instead.

In *Reasons and Persons* and *On What Matters*, Derek Parfit argues that it is a constraint on any plausible moral theory that morality must *never* be directly collectively self-defeating, and he claims that the most plausible versions of consequentialism, contractualism, and Kantian ethics all imply that morality is never directly collectively self-defeating:

> …moral principles or theories are intended to answer questions about what *all* of us ought to do. So such principles or theories clearly fail, and condemn themselves, when they are directly self-defeating at the collective level. (2011, 306, italics in the original)

> [The assumption that morality is never directly collectively self-defeating] is either made or implied by most of the many different theories [of morality]. (1984, 113)

Some theorists not only agree with Parfit that *morality* can never be directly collectively self-defeating, but also believe that *rationality* and other forms of normativity can never have that property either. For example, the Kantian idea that our acts or principles must be willable as universal law might be taken as a way of suggesting that both morality and rationality can never be directly collectively self-defeating. And even theorists who grant that there is a narrow form of rationality that is sometimes directly collectively self-defeating often insist that there is a broader and more important form of rationality, sometimes called 'enlightened self-interest', that never has that property.

These theorists are all mistaken, because morality and all other interesting forms of normativity are sometimes directly collectively self-defeating. To see why, consider cases like the following:

> **Stampede Case**
> We find ourselves in an enormous stampede. Unless everyone immediately stops stampeding, it is clear that some of us will be moderately harmed. However, it is also clear that everyone will not immediately stop stampeding, and so anyone who does stop stampeding will be severely harmed in a way that does no good for anyone else.

In this case, from every interesting normative perspective – self-interest, enlightened self-interest, morality, benevolence, and so on – each person is *required* to continue stampeding, despite the fact that it is clear that the outcome would be *better* in every normatively interesting sense if everyone did not continue stampeding instead. This shows that morality and all other forms of normativity are sometimes directly collectively self-defeating, because it shows that there are cases in which everyone can be sure that if each person does what is required, the result will be worse from the perspective of each than if each had not done what is required instead.

Here is another example:

> **Units of Good Case**
> 1,000 people are put into isolation booths. It is common knowledge that each must choose between Options A and B, with the following outcomes: If everyone chooses A, then each receives 99 additional units of good; if everyone chooses B, then each receives 100 additional units of good; otherwise, each person who chooses A receives 10 additional units of good and each person who chooses B loses a catastrophic 100,000 units of good.

In this case, from every interesting normative perspective, each person is *required* to choose A, despite the fact that it is clear that the outcome would be *better* from each person's perspective if everyone did not choose A instead. Once again, this shows that morality and all other forms of normativity are sometimes directly collectively self-defeating, because everyone can be sure that if each does what is required, the result will be worse than if each had not done what is required instead. Importantly, these conclusions follow even in cases such as these that do not involve any uncertainty or normative failure, in which it is common knowledge that: everyone will satisfy their normative requirements, everyone knows the relevant facts, and everyone knows which course of action would lead to the best outcome.[3] This ensures that these cases are genuine counterexamples to the thesis that morality is never directly collectively self-defeating in the sense intended by Parfit and others.

Why are morality and other forms of normativity directly collectively self-defeating in these cases? The answer is that a particular form of risk aversion is sometimes required: in particular, sometimes even when it is common knowledge that everyone will satisfy their requirements and that everyone is fully informed, it is also clear that

[3] On the natural and intended understanding of these cases, there is a sense in which everyone has the same options, one of which is such that each individual knows enough about what the others will do to know of that option that the outcome will be objectively best if s/he choose that option, and thus there is a particular option such that each knows that s/he will successfully follow morality only if s/he chooses that option. In this way, these cases are not *unsettled coordination problems* in which it is unknown which course of action would lead to the best outcome. (Compare the unsettled coordination problems discussed by Parfit, 1984, 53–4.)

the option that would lead to the best outcome if universally chosen is associated in a way that is salient to everyone with great risks without compensating rewards, and in some such cases each person can, by this very reasoning, know that others will coordinate on a 'risk-averse' option instead, thereby ensuring that each person is required to choose that 'risk-averse' option themselves, even if it is clear that everyone choosing that risk-averse option guarantees a worse outcome from the perspective of each than if everyone did not choose that option instead.[4] In the words of David Lewis in another context, in these cases individuals can be seen as reaching "a coordination equilibrium that is somehow salient: one that stands out from the rest by its uniqueness is some conspicuous respect. It does not have to be uniquely *good*; indeed, it could be uniquely bad. It merely has to be unique in some way the subjects will notice, expect each other to notice, and so on" (1969, 35, italics in the original).

In addition to showing that all forms of normativity are sometimes directly collectively self-defeating, the preceding considerations also show that an important research program on morality and game theory is misguided, because the essential and guiding assumption of that research program is that morality always guarantees optimal cooperation when it is common knowledge that: everyone has full information about the symmetrical choices facing everyone, will act freely, will satisfy their normative requirements including moral requirements, and knows of a unique option that the outcome would be best if that option were chosen by everyone.[5]

Having argued that morality and all other forms of normativity are sometimes directly collectively self-defeating (DCSD), it is useful to consider further implications for moral theory.

First, consider the Kantian idea that an act is permissible only if the maxim behind that act is willable as universal law. What does this mean? Suppose one does not know exactly what this means. Nonetheless, one can know on the basis of the arguments above that if this implied that morality is never DCSD, then it would be false. More generally, consider versions of 'Kantian ethics', 'rule utilitarianism', 'utilitarian generalization', 'cooperative utilitarianism', or any other view on which a notion of 'universalizability' seems to play a guiding role. Because we can show that morality is sometimes DCSD, we can show that such views would be false if they implied that morality is never DCSD. As a result, we should not interpret such views as having that implication – contrary to Parfit's claims – if we want to develop the most plausible versions of these views.

Recognizing that morality is sometimes directly collectively self-defeating might

[4] Note that this 'risk aversion' in this sense does not imply departure from standard decision theory.

[5] For an brief description of this research program, see Parfit, 1986, pg. 867; for more detail, see Regan, 1980, especially pp. ix –xi, and pp. 4–5, and Gibbard, 1971, especially pp. 6–9.

also lead us to reexamine beliefs about what individuals are required to do in real-world collective action problems. For example, consider the following:

Pollution Case
Each of us will do better by not reducing emissions than by reducing emissions; however, at the same time, each of us will do substantially worse if no one reduces emissions than we would if everyone reduced emissions.

Many would say that each of us is required to reduce emissions in this case because the alternative is directly collectively self-defeating. However, that is a bad argument, because morality and all other forms of normativity are sometimes DCSD. So, if individuals are required to reduce emissions in such a case, it must be for some other reason, such as the impermissibility of the *harm* that is done by those emissions.

In response to all of this, it might be claimed that although morality is sometimes *mildly* DCSD as in the Stampede Case and the Units of Good Case above, it can never be *dramatically* DCSD.

At first glance, this response might seem promising. However, it does not succeed, because morality and all other forms of normativity are sometimes dramatically directly collectively self-defeating. To see why, consider cases like the following:

Dramatic Stampede Case
We find ourselves in an enormous stampede. Unless everyone stops stampeding, it is clear that an increasing number of people will be seriously harmed and killed. However, it is also clear that everyone will not stop stampeding, and so anyone who does stop stampeding will be severely harmed or killed in a way that does no good for anyone else and simply adds to the ultimate aggregate harm caused by the stampede.

This case is representative of real-life stampedes. In such cases, individuals are not required to stop stampeding, even if continuing is dramatically directly collectively self-defeating.

Here is an infinitely dramatic example:

One Million Dollars Case
Everyone on the planet is isolated and instructed to choose a number, and a neon sign reading 'One Million' is lowered in front of each person. If everyone chooses the same number, then the standard of living of each person in the world will be increased by an amount equivalent to one one-millionth of that number of dollars; otherwise, if everyone fails to choose the same number, each person's standard of living will be dramatically reduced. All of this is common knowledge.

What number should each choose in this case? Each should choose one million, because it is common knowledge that one million is uniquely salient to everyone, which makes it common knowledge that one million is the only number that has any chance of being chosen by everyone, which makes it the case that each should choose that number, given the dramatic costs of a failure to coordinate. However, if everyone chooses one million, the standard of living of each person in the world will remain the same rather than rising by, say, $1 billion each, which it is clear that everyone could bring about by simply by choosing the number one quadrillion instead of one million (and so on for any amount whatsoever). As this shows, morality and all other forms of normativity are sometimes *catastrophically directly collectively suboptimal*, because they sometimes direct everyone to choose an option that is certain to lead to a catastrophically worse outcome than an antecedently identifiable option that they could have directed everyone to choose instead. This is truly catastrophic, because instead of solving all of the world's material problems, following morality and other forms of normativity in such a case would not do anyone any good at all.[6]

Here is another dramatic example:

> **End of the World Case**
> Aliens come to Earth and force each family on the planet to choose between 'cooperating' and 'defecting', which are known to have the following consequences: If all choose to cooperate, the aliens will leave and everyone's life will go on the same as before – but if even one family chooses to defect, in one year the aliens will destroy the Earth and every living thing on the Earth, and in the meantime will ensure that each family that chooses to cooperate has a miserable life of intense suffering, while each family that chooses to defect has a wonderful and flourishing final year on Earth.

A philosopher might insist that every family would be required to cooperate in this case. But upon reflection, it is clear that if a billion families were actually in this situation, then some of them, by this very reasoning and without making any moral mistake, would choose to defect, thereby ensuring the end of the world in one year, and ensuring that cooperation would mean a futile sacrifice of one's own family in a way that was impermissible. This illustrates the way in which morality can be *catastrophically* directly collectively self-defeating.

In response to all of this, a philosopher might insist that it is simply *absurd* to think that morality is sometimes dramatically directly collectively self-defeating. Such a

[6] This example is based on a case discussed by Schelling, 1957, who also introduces the relevant notion of *salience*. For ease of exposition, it is assumed here that one dollar would not make a difference to anyone on the planet.

thought is true in an important sense – it is true in the same sense that we sometimes find ourselves in situations that are absurd. But *absurdity* in that sense does not give rise to a reductio – just as finding ourselves in an *absurd* situation such as the End of the World Case would not show that we were not in that situation.

What does contractualism say about all of this? It is unclear. Contractualism is, roughly, the view that an act is required if it is required by principles that we would agree upon in some special scenario, or, alternatively, if it is required by principles that we could not reasonably reject. One problem for contractualism is that everyone would want everyone to cooperate in the End of the World Case, and everyone would agree to cooperate if such agreement was possible and binding; but if it were supposed to follow from these facts that contractualism implies that individuals are required to cooperate in the End of the World Case, then the view would be false, and would be false because it ignores every interesting aspect of collective action problems.

In response, contractualists would quite reasonably insist that their view does not mistakenly imply that cooperation is required in the Dramatic Stampede Case and the End of the World Case. But if that is correct, then they should also be quick to admit that their view does not give us any easy answers about how to think about challenging collective action problems such as those discussed above in the way it would if such a view were never DCSD. Similar remarks apply to universalization theories: either those theories are false because they imply that morality is never DCSD, or else they do not provide any immediate guidance about how to think about such collective action problems.[7]

The upshot is that morality and all other forms of normativity are sometimes dramatically directly collectively self-defeating, which means that many influential normative theories are either false, or at least don't have the consequences that their adherents take them to have.[8] One important consequence is that morality and other forms of normativity cannot be relied upon to solve collective action problems even in a world of normatively flawless agents. In particular, even if a disaster will ensue if everyone acts in a particular way or on a particular principle, that does not settle the

[7] The cases above are counterexamples even to sophisticated never directly collectively self-defeating universalizability theories that are intended to apply to non-ideal situations. For example, consider Parfit's principle "Follow the rules whose being followed by everyone would make things go best, unless some other people have not followed these rules, in which case do whatever, given the acts of others, would make things go best" (Parfit, 2011, pg. 317). This principle seems to deliver the mistaken verdict that everyone is required to choose B in the Units of Good Case, because at the moment that everyone chooses in that case, no one has yet failed to follow optimific principles, and so the principle seems to imply that each must choose B. Similar remarks apply regarding most of the other cases above.

[8] It is worth noting that the arguments here do not depend on any controversial premises. In particular, the arguments here do not depend on the controversial premise that the *better than* relation is intransitive; compare Rachels, 1989, and Temkin, 2011. Even if the *better than* relation were intransitive, that would not show that morality is sometimes *dramatically* DCSD, as the arguments here reveal.

question of whether individuals are permitted to act in that way or on that principle. And because many of the most important questions about modern moral life are essentially questions about what individuals are required to do in such situations – for example, what individuals are required to do about climate change, what individuals are required to do when products are produced in morally objectionable ways – an important practical upshot is that such questions cannot be answered by asking 'But what if everyone did that?', or by more sophisticated appeals to 'universalizability'.

In response to the preceding arguments, theorists who are focused on Parfit's writings sometimes object that the notion of direct collective self-defeat is merely a technical notion introduced by Parfit, and as a result it is impossible to evaluate the arguments above without examining Parfit's precise definition.[9]

This objection is misguided, because as Parfit's own discussion makes clear, direct collective self-defeat is *not* a technical notion even on his view. For example, when Parfit first discusses that notion in *Reasons and Persons*, he assumes that we can all grasp that notion independent of any definition by reflecting on prisoner's dilemmas and other social dilemmas, where these examples illustrate the importance of that intuitive notion to normative theory. He then considers a provisional definition that might initially seem to capture that notion, and then immediately rejects that provisional definition. Why? Because Parfit argues that when we consider a hypothetical case, we can see that the provisional definition gives a different verdict than the intuitive notion that we care about; thus, the provisional definition must be rejected.[10] This shows that the notion of direct collective self-defeat is not a technical notion even for Parfit. Instead, Parfit correctly recognizes that direct collective self-defeat is an intuitive notion that we can all grasp on the basis of reflection on social dilemmas and independent of any stipulative definition, and that this non-technical notion is of central importance to normative theory – and when we consider how this non-technical notion applies to the cases presented above, we see that morality and all other forms of normativity are sometimes directly collectively self-defeating. (This and other issues related to Parfit's views are discussed in more detail below.)

Deontologists might object that this entire discussion depends on a sense of *betterness* that is foreign to their view, because (they might say) their view is concerned with *acts* rather than *outcomes*.[11] However, such an objection is misguided. If we all

[9] In conversation.

[10] Parfit, 1984, pp. 53–54. Unlike the coordination problems that Parfit uses in these passages to reject the provisional definition of DCSD, the examples in this paper are not merely cases where morality *fails to direct us toward* the morally best outcomes, but are cases where morality *directs us away from* the morally best outcomes, thereby constituting cases in which morality is genuinely DCSD (Parfit, 1984, pg. 54).

[11] For such an objection, see Adams, 1997, pg. 259.

continue stampeding in the Stampede Case, it is certain that we will cause harm, whereas if we all stop stampeding, it is certain that we will do no harm at all. As a result, it is perfectly sensible and correct to say that, collectively, continuing stampeding is deontologically worse than stopping stampeding, but that nonetheless each of us individually is required to continue stampeding, because if an individual were to stop, s/he would do something (namely, severely harm an innocent person – him or herself) that is deontologically worse than what s/he would do by continuing stampeding. That is why in the Stampede Case deontology is DCSD.[12]

More fundamentally, some deontologists might reject the judgments about cases appealed to above, and insist instead that, for example, one has a moral obligation to stop and be trampled to death in the stampede case even though doing so would do no good for anyone. In light of this possibility, the arguments above should officially be understood as conditional on the judgments appealed to above. If one accepts those judgments – as many theorists and almost all ordinary people do – then the conclusions above about direct collective self-defeat follow; if one rejects those judgments, then these arguments still establish an interesting conditional result, the consequent of which can be resisted only by endorsing verdicts on cases that many find highly counterintuitive.

A more subtle objection comes from agent-neutral consequentialists, some of whom believe that it is a clear and important virtue of their view that it is never directly collectively self-defeating. For example, Parfit argues:

> [Agent-neutral consequentialist theories] cannot be directly self-defeating, since [they are] *agent-neutral*: giving to all agents *common* moral aims. (1984, 54–55, italics in the original)[13]

> …Common-Sense Morality is often directly collectively self-defeating. [But] a moral theory must be collectively successful. [Those who believe in Common-Sense Morality] must therefore revise their beliefs, moving from [Common-Sense Morality to a form of agent-neutral consequentialism]. (1984, 111)[14]

This is Parfit's main argument for the sort of moral theory he favors in *Reasons and Persons*. Unfortunately, this argument is unsound, because agent-neutral consequentialism is sometimes directly collectively self-defeating, as illustrated by the Units of

[12] Another example from Parfit: "Suppose that each could either (1) carry out some of his own duties or (2) enable others to carry out more of theirs. If all rather than none give priority to their own duties, each may be able to carry out fewer. Deontologists can face [situations in which their theory is DCSD]" (Parfit, 1984, pg. 98).

[13] Parfit, 1984, pp. 54–55, italics in the original.

[14] See also Parfit, 2011, pg. 306.

Good Case above.[15] For a further example that may be useful in anticipating various avenues of reply, consider the following complicated variant of the Units of Good Case, and the graph that follows, which represents the possible outcomes in this more complicated case:

Complicated Units of Good Case
1,000 people are put into isolation booths, and each must choose between Options A and B, with the following outcomes: If everyone chooses A, then each receives 99 additional units of good; if everyone chooses B, then each receives 100 additional units of good; otherwise, every person who chooses A receives 10 additional units of good and every person who chooses B loses a catastrophic 100,000 units of good, unless one and only one person chooses A, in which case every person receives 101 additional units of good. All of this is common knowledge, as is the fact that everyone will successfully follow agent-neutral consequentialism.

The graph in Figure 1 is a simplified representation of the possible outcomes in this case:

Figure 1. Complicated Units of Good Case.

[15] Rabinowicz, 1989 argues that *some* versions of agent-neutral conquentialism can be directly collectively self-defeating. My argument aims to show that *all* plausible versions of agent-neutral consequentialism are sometimes directly collectively self-defeating.

In this case, it is common knowledge that everyone will satisfy their requirements, that everyone is fully informed, and that everyone can see that the option that would lead to the best outcome if universally chosen (B) is associated in a way that is salient to everyone with great risks without compensating rewards; as a result, each can be certain that the others will tend to coordinate on the risk-averse option A, which ensures that each person is *required* to choose that risk-averse option A themselves (because they know that their choice would otherwise make the outcome worse on agent-neutral consequentialist grounds), even though it is clear that everyone choosing that risk-averse option A guarantees a worse outcome from the perspective of each (bringing about the yellow dot outcome on the graph) than if everyone did not choose that option instead (bringing about the red dot outcome on the graph). As a result, agent-neutral consequentialism is directly collectively self-defeating in this case because everyone can be certain that: *if we all successfully follow agent-neutral consequentialism by each doing what is actually required, we will thereby cause our agent-neutral consequentialist aims to be worse achieved than they would have been if none of us had done what is actually required*.[16] In more detail: because of what each knows about the situation and thus what each knows about how the others will choose, each can be certain that choosing A will make the outcome objectively better than choosing B, and thus agent-neutral consequentialism requires each to choose A; at the same time, each can be certain that if each failed to do what is required and therefore chose B, the outcome would be better even though no one would then satisfy agent-neutral consequentialism (because for each it would be true that there is something else s/he could have done (namely, choose A) that would have led to more good (by bringing about the blue dot outcome). Thus, agent-neutral consequentialism is sometimes directly collectively self-defeating.

Is this a bad result for agent-neutral consequentialism? No. It would be a bad result for agent-neutral consequentialism if it were never DCSD, because we've seen that all plausible normative theories are sometimes DCSD.

Why then does Parfit think that agent-neutral consequentialism is never DCSD? Parfit offers the following sufficient conditions for direct collective self-defeat:

> A theory T is directly collectively self-defeating when:
> (i) it is *certain* that, if we all successfully follow T, we will thereby cause our T-given aims to be worse achieved than they would have been if none of us had successfully followed T, or

[16] Compare (i) on page 54 of Parfit, 1984.

(ii) our acts will cause our T-given aims to be best achieved only if we do not successfully follow T.

Based on these conditions, Parfit offers the following argument that agent-neutral consequentialism is never DCSD:

> [Agent-neutral consequentialism] cannot be directly self-defeating, since it is *agent-neutral*: giving to all agents *common* moral aims. If we cause these common aims to be best achieved, we must be successfully following this theory. Since this is so, it cannot be true that we will cause these aims to be best achieved only if we do not follow this theory. (1984, 54—55)

In the last sentence of the preceding quote, Parfit concludes that it is necessarily false that: our acts will cause our agent-neutral consequentialist aims to be best achieved only if we do not successfully follow agent-neutral consequentialism, which is an instance of (ii), where 'agent-neutral consequentialism' replaces 'T'. From this, it is supposed to follow that agent-neutral consequentialism is never DCSD.

At this point, someone might object to Parfit's argument as follows: "On Parfit's analysis, a theory can be DCSD in either way (i) or way (ii), and Parfit has shown only that agent-neutral consequentialism cannot be DCSD in way (ii); so, it doesn't follow from Parfit's premises that agent-neutral consequentialism cannot be DCSD in way (i), and so it doesn't follow that agent-neutral consequentialism is never DCSD."

In reply to this objection, Parfit would presumably insist that (i) is to be understood in such a way that (i) implies (ii). If that's right, then Parfit's demonstration that agent-neutral consequentialism can never be DCSD in sense (ii) also shows that it can never be DCSD in sense (i).

However, even granting such a reply, Parfit's argument still faces a decisive objection. To see the problem, note that even if (i) implies (ii), Parfit's argument is still invalid as stated:

> If C is ever (i) or (ii), then C is sometimes DCSD.
>
> C is never (ii).
>
> Therefore, C is never (i), since (i) implies (ii).
>
> Therefore, C is never DCSD.

If the problem is not immediately apparent, it might help to combine the two middle claims:

If C is ever (i) or (ii), then C is sometimes DCSD.

C is never (i) or (ii).

Therefore, it is not the case that C is sometimes DCSD.

This argument is invalid because it denies the antecedent. To get a valid argument, we would have to understand the first premise as a biconditional, and thus we would have to interpret (i) and (ii) as together yielding a full analysis of direct collective self-defeat. However, Parfit explicitly claims that (i) and (ii) provide only sufficient conditions for direct collective self-defeat, and not a full analysis.[17] As a result, Parfit's argument is invalid, because it has the invalid form above.

Of course, this raises the question of whether (i) and (ii) can in fact yield a full analysis of direct collective self-defeat – in other words, it raises the question of whether the following is true:

A theory T is DCSD when *and only when* either (i) is true or (ii) is true, where (i) and (ii) are understood in the way that Parfit intends.

This *Implicit Analysis* is false, because it does not capture the essence of direct collective self-defeat, including the essential idea that a theory is DCSD when it *directs us toward outcomes that are certain to be worse* (1984, 54). In particular, the Implicit Analysis fails to deliver the correct verdict on the cases discussed above in which:

(iii) it is common knowledge that: everyone knows the relevant facts, will act freely, will satisfy their normative requirements, and everyone can also be certain that: if each does what T actually requires, the T-given aims of each will be worse achieved than they would have been if none had done what T actually requires.

At the very least, such cases show that (iii) is an additional sufficient condition for direct collective self-defeat, which means that Parfit's argument that agent-neutral consequentialism is never DCSD cannot be salvaged, because (iii) together with the units of good cases show that agent-neutral consequentialism is sometimes DCSD.

In response, a defender of the Implicit Analysis might say "But consider the possibility that in the Complicated Units of Good Case one and only one player chooses Option A; then, each person successfully follows agent-neutral consequentialism and brings about the best outcome; this shows that even in the Complicated Units of

[17] Parfit makes this explicit in the following passage, where he explains how he intends the phrase "[A theory T is] directly collectively self-defeating when..." to be understood: "By 'when' I do not mean 'only when'" (1984, 54).

Good Case agent-neutral consequentialism does not direct us toward outcomes that are certain to be worse."

This reply gives the phrases 'direct us toward' and 'successfully follows' a meaning that is very different from their meaning in the intuitive thought that a theory is DCSD when it directs us toward outcomes that are certain to be worse, or when it is certain that the outcome would be worse if each successfully followed the theory than if each did not. More specifically, this reply involves a backward-looking conception of *directing an agent toward an outcome* and *successfully following a theory* that is irrelevant to any interesting normative concept. To see why, return to the players in the Units of Good Case and assume that all the players evaluate options and make their decisions simultaneously as well as independently. Now consider the point in time as they are about to make their decisions. At that point in time, does agent-neutral consequentialism direct the players toward a particular outcome? It does in every intuitive sense – namely, the outcome in which everyone chooses Option A: after all, even before anyone chooses Option A, each player *knows* that choosing Option A will lead to an objectively better outcome, and thus agent-neutral consequentialism directs each player to choose Option A, and thus each player successfully follows agent-neutral consequentialism only if that player chooses Option A.

The Implicit Analysis denies all of this. Instead, on that analysis agent-neutral consequentialism gives the players no direction at all before their decisions are made, on the grounds that there are multiple combinations of choices that would result in satisfaction of agent-neutral consequentialism. That is how the Implicit Analysis insists that agent-neutral consequentialism does not direct the players away from the best outcome: according to the analysis, there are no facts about what agent-neutral consequentialism directs the players to do until after everyone has made their decision, at which point the theory 'directs' everyone to have chosen in such a way that they now each satisfy agent-neutral consequentialism. However, this is a revisionary account of how agent-neutral consequentialism directs agents toward outcomes – because it entails, contrary to the claims of all actual consequentialists, that what agents know about the consequences of their choices has no relevance to what consequentialism directs them to do – and more importantly it is also an unacceptable account, because any interesting normative theory must provide direction for our decisions, and not only after they are made.

In response, a defender of the Implicit Analysis could attempt to bite the bullet and simply insist that agent-neutral consequentialism offers no direction in such cases until after decisions are made. However, the costs of such a stance prove unacceptably high when applied to other cases, especially cases that involve physical indeterminacy with no residual epistemic uncertainty. For example, consider a case that is similar to

the Units of Good Case, but where the uncertainty of the outcomes derives entirely from physical indeterminacy:

> **One-Player Units of Good Case**
> You know that you alone must choose between the following two options, and that your choices will have the following consequences for yourself and 999 other innocent people:
>
> Option A: 99% chance that everyone receives 99 additional units of good; 1% chance that everyone receives 10 additional units of good.
>
> Option B: 1% chance that everyone receives 100 additional units of good; 99% chance that everyone receives negative 100,000 units of good.

Suppose that the chances in this case are purely physical and that there is no residual epistemic uncertainty. (For example, suppose that physicists have designed a non-deterministic pleasure and pain dispensing device to have these properties; you will simply choose whether to press the 'Option A' or 'Option B' button.)

On any sensible interpretation, agent-neutral consequentialism directs you to choose Option A in this case, which means that you successfully follow agent-neutral consequentialism only if you choose Option A. Would defenders of the Implicit Analysis agree? If they do not, then they are committed to the view that agent-neutral consequentialism never provides any actual guidance to our decisions, because physical indeterminacy always underlies all of our decisions. So, to avoid this result, they would presumably agree that agent-neutral consequentialism directs you to choose Option A in this case.

But if that is right, then there is a powerful argument that agent-neutral consequentialism directs each player to choose Option A in the original Units of Good Case. For consider that, for each player in that original case, there is some distribution of credences that that player ought to have, given his or her evidence, about how the other players will choose. Given that distribution of rational credences, we can imagine a one-player game with the same outcomes and probabilistic structure, but where the probabilities arise from physical indeterminacy with no residual epistemic uncertainty as in the One-Player Units of Good Case. If, as we are assuming, agent-neutral consequentialism directs you to choose Option A in the One-Player Units of Good Case, then it also directs each player to choose Option A in the one-player game that is derived in such a way from his or her rational credences in the original Units of Good Case. But if agent-neutral consequentialism directs each player to choose Option A in the one-player games that are derived from their rational credences, then

it also directs each player to choose Option A in the original Units of Good Case itself, because there is no normatively relevant difference between the choices that each individual would face in those one-player games and the corresponding choices that they face in the original Units of Good Case. As a result, initial defenders of the Implicit Analysis are forced either to abandon that analysis by admitting that agent-neutral consequentialism directs players to choose Option A in the original Units of Good Case, or else to bite an unacceptable bullet and insist that agent-neutral consequentialism almost never provides any guidance to our decisions at all, because physical indeterminacy always underlies our decisions.

The preceding discussion shows the importance of distinguishing between a normative theory's theory of objective value and its theory of choice. As we have just seen, a theory of objective value never directs us toward any outcomes itself – it is only in conjunction with a theory of choice that we are directed to make particular choices, and thereby directed toward particular outcomes. This shows that the analysis of direct collective self-defeat under consideration must be inadequate, because that analysis focuses only on satisfaction of a theory's theory of objective value, and not on satisfaction of its theory of choice. In other words, that analysis must be inadequate because the notion of direct collective self-defeat is about what a theory directs us toward, and without a theory of choice a theory never directs us toward anything at all.

For these reasons, an adequate full analysis of direct collectively self-defeat must be tied to a normative theory's theory of choice. Here is a proposal:

> A theory T is directly collectively self-defeating (DCSD) when: it is certain from the perspective of each of us that, if each of us successfully follows T's theory of choice, we will thereby cause our T-given aims to be worse achieved than they would have been if none of us successfully followed T's theory of choice.

If successfully following T's theory of choice is the same as doing what T requires, then this *New Analysis* is equivalent to the claim that: *A theory T is directly collectively self-defeating (DCSD) when: (from the perspective of each of us) it is certain that, if each of us does what T requires, we will thereby cause our T-given aims to be worse achieved than they would have been if none of us did what T requires*. This amounts to an analysis of direct collective self-defeat for cases in which everyone has the same two options. To test this analysis, we can consult our judgments about cases, and our judgments about the concept of direct collective self-defeat. Upon reflection, this New Analysis delivers the correct verdict on all of the cases that theorists have discussed in connection with direct collective self-defeat, and, unlike the Implicit Analysis discussed above, also fits our intuitive concept of direct collective self-defeat, according to which a theory is

DCSD when it directs each of us to act in a way that is certain to be worse than if everyone did not follow the theory's directions instead.

In response, a defender of the Implicit Analysis might raise the following objection: "Perhaps the New Analysis and/or (iii) captures the idea that a theory is DCSD when it directs us toward outcomes that are certain to be worse. But that idea is inconsistent with other more firmly held beliefs that we have about self-defeat, and so the notion of direct collective self-defeat must be regimented in a different way – most likely, in the way the Implicit Analysis suggests. That is because Donald Regan, Derek Parfit, and others have provided cases that show that normative theories sometimes direct us away from the best outcomes, but are not thereby self-defeating." What the objector has in mind are cases like the following:

> **Miners Case**
> Suppose that several miners are trapped, with floodwaters rising. Before we can find out where these miners are, we must decide which floodgate to close.

The possible outcomes of our decision are outlined in Table 1.

Table 1. Miners Case

	The miners are in Shaft A	The miners are in Shaft B
We close Gate 1	We save ten	All die
We close Gate 2	All die	We save ten
We close Gate 3	We save nine	We save nine

Assume that, on the evidence, the miners are equally likely to be in either shaft.[18]

In this case, we are required to close Gate 3, even though it is certain that we will thereby bring about an outcome that is not best; nonetheless, this does not show that normativity is directly collectively self-defeating. Does this undermine the idea that (iii) is a sufficient for direct collective self-defeat?

It does not. What the Miners Case shows is that there is a crucial distinction between, on the one hand, *it being certain that an option will lead to an outcome that is not best* and, on the other hand, *it being certain that an option will lead to a worse outcome than some other antecedently identifiable particular option*, and that a theory is DCSD when it directs us to choose an option of the latter type, but not when, as in the Miners Case, it merely directs us to choose an option that is certain to be not best. In particular, if we close Gate 3 we bring about an outcome that is certain to be not best, but we do not bring about an outcome that is certain to be worse than the outcome

[18] This example is taken from Parfit, 1988, pp. 2-3, who follows Regan, 1980, pg. 265.

of any particular other option, because there is no other option that is antecedently *certain* to lead to a better outcome than closing Gate 3. This is in perfect tune with conditions (i), (ii), and (iii) above, because the natural way of extending those conditions to cases involving many options such as the Miners Case is by claiming that a theory is DCSD when it directs everyone to choose an option that is *certain* to lead to a worse outcome than an antecedently identifiable alternative option that it could have directed everyone to choose instead – but not when, as in the Miners Case, the theory merely directs everyone to choose an option that is certain to lead to an outcome that is not best.[19] As a result, the Miners Case does not ultimately raise a problem for the intuitive notion of direct collective self-defeat, and does not raise a problem for the view that conditions (i), (ii), and (iii) are each sufficient for direct collective self-defeat, and does not raise a problem for the New Analysis above.

In response to all of the preceding arguments, it might be objected that theories like agent-neutral consequentialism still imply that it is always metaphysically possible to bring about the outcome that is best without anyone acting in a way that is wrong – and that such a possibility of doing what is best without anyone doing wrong is how the notion of direct collective self-defeat is best understood. However, although optimal cooperative action is *metaphysically possible* in cases such as the units of good cases, each individual is also certain that such cooperation will not obtain, and as a result from the perspective of each individual *it is certain that the aims of morality will be worse achieved if each successfully follows morality than if everyone did not successfully follow morality instead* – which is of course just to say that morality is directly collectively self-defeating, because it would actually be wrong to act in accord with optimal cooperative action based on the full information of the case and what each knows about the morally flawless dispositions of others. This shows that agent-neutral consequentialism ultimately has no interesting advantage over other types of ethical theories with respect to direct collective self-defeat.

The arguments above also cannot be dismissed by simply insisting on an alternative definition of direct collective self-defeat on which the arguments above do not go through – for example, a stipulative definition on which (iii) is not a sufficient condition for direct collective self-defeat. In part, this is because direct collective self-defeat, like knowledge, is a notion that we track and care about prior to seeing any stipulative definition, as is illustrated by our interest in social dilemmas and other situations in

[19] Such an extension presumably must be restricted to cases in which everyone chooses between the 'same' options, where those options are individuated in a 'natural' way – and in other cases the notion of direct collective self-defeat seems to have no clear application. The Miners Case could also be redescribed as a two-option case, where Option One is to close Gate 3, and Option Two is to close one of the other gates. An analysis that includes the features advocated here also delivers the correct verdict given that description, because choosing Option One is *not certain* to lead to a worse outcome than choosing Option Two, and therefore such an analysis does not imply that morality is sometimes DCSD.

which self-interest is directly collectively self-defeating, and so direct collective self-defeat is not a notion that we are free to define however we like if the result is to have any interest to normative theory. More specifically, insofar as we should care whether a theory is sometimes directly collectively self-defeating, that is because having that property means that regrettable consequences are assured even in cases like those described in (iii) in which it is common knowledge that everyone knows the relevant facts and will successfully follow the theory. As a result, a definition on which satisfaction of (iii) is not sufficient for direct collective self-defeat has no practical or theoretical interest, not only because it does not track the important notion of 'directing us toward outcomes that are certain to be worse', but more importantly because it does not track the kind of collective self-defeat that it is regrettable for a theory to imply – because the most regrettable form of collective self-defeat is when a theory is collectively self-defeating in the sense of (iii), when it is collectively self-defeating even though it is common knowledge that everyone knows the relevant facts and will do what is required, and that regrettability is not mitigated in any interesting way when it is also true that if individuals had failed to do what they actually know they are required to do, the outcome could have been better. As a result, any discussion that rejects (iii) as a sufficient condition for direct collective self-defeat is doomed to reduce to a definitional exercise that has no connection to any property that we should care whether a normative theory has – whereas endorsing (iii) is essential to capturing the kind of collective self-defeat that is of central interest from both a practical and theoretical perspective.[20] So, such a stipulative definition has no chance of playing an interesting role in arguments about the nature of morality, and in particular has no chance of playing an interesting role in arguments against commonsense morality.

The preceding discussion suggests the following evaluation of Parfit's main argument for never-directly-collectively-self-defeating moral theories:

Parfit's Main Argument
To be plausible, a moral theory must be never DCSD.
So, we must reject common-sense morality and other theories that are sometimes DCSD, and instead endorse a version of never-DCSD moral theory.

The premise is false, because morality is sometimes DCSD. As a result, not only is it consistent to deny the conclusion, but there is decisive reason for thinking that the conclusion is false, because a moral theory is false if it is never DCSD. If we were to

[20] An additional consideration is that many theorists, including Parfit, take facts about what would be wrong when agents know the relevant facts as explanatorily fundamental – which provides decisive reason to think that the sense of direct collective self-defeat captured by (iii) is the sense that must have the greatest theoretical importance, because (iii) is explicitly concerned with whether a theory is collectively self-defeating when everyone knows the relevant facts. (See Parfit, 2011, Section 21.)

follow Parfit in thinking that consequentialists, contractualists, Kantian theorists, and most others have been "climbing the same mountain" toward the goal of developing the most plausible version of never-DCSD moral theory, then this would mean that those theorists have all been climbing the wrong mountain.[21]

This is not to denigrate Parfit's work, which has the highest virtues of clarity, testability, originality, and importance. Because Parfit's work has such virtues, identifying a clear objection to his arguments leads to important progress in normative theory.

In sum, morality and all other interesting forms of normativity are sometimes dramatically directly collectively self-defeating, which means that many influential normative theories are either false, or at least don't have the consequences that their adherents take them to have. In particular, morality and other forms of normativity cannot be relied upon to solve collective action problems even in a world of normatively flawless agents. A practical upshot is that many of the most important questions about modern moral life cannot be answered by asking 'But what if everyone did that?', or by a more sophisticated appeal to a form of 'universalizability'.

Appendix: The Equilibrium Objection

In "Group Morality", Frank Jackson uses an example that bears some similarity to the Stampede Case to argue that it is possible to "have a group action which is wrong, yet every constituent act is right; and a group action which is right yet every constituent act is wrong" (1987, 102). Parfit accepts Jackson's conclusions in later work, but neither Parfit nor Jackson take these conclusions to show that morality is sometimes directly collectively self-defeating.[22] This appendix shows that Jackson's conclusions do not clearly follow from the example he discusses, and that his discussion cannot be extended to show that morality is sometimes directly collectively self-defeating (DCSD) – but that such conclusions are vindicated by the examples discussed above, despite an important objection that is suggested by reflection on Jackson's discussion.

[21] For this metaphor and a summary of Parfit's arguments that the most plausible versions of consequentialism, contractualism, and Kantian ethics all imply that morality is never DCSD, see Parfit (2011, 25-26). Parfit endorses Parfit's Main Argument in Parfit (2011, 306) : "In [social dilemmas], in acting on common sense moral principles, we are acting in ways that are directly collectively self-defeating. If we were Rational Egoists, that would be no objection to our view, since this form of Egoism is a theory about individual rationality and reasons. But moral principles or theories are intended to answer questions about what all of us ought to do. So such principles or theories clearly fail, and condemn themselves, when they are directly self-defeating at the collective level". See Parfit (2011, 111 and 113) for an earlier discussion and more explicit presentation of the argument.

[22] Jackson does not claim that his conclusions show that morality is sometimes DCSD, and in later work Parfit continues to rely on the premise that morality is never DCSD despite Parfit's endorsement of Jackson's conclusions in Parfit, 1988.

Here is Jackson's example:

> [Suppose that] There is a steady stream of traffic going to work. Everyone is driving at 80 kilometres per hour. It would be safer if everyone was driving at 60. The right group action is for everyone together to drive at 60. But what about each person, should he or she drive at 60? The answer may well be no; for it may well be the case that if he or she were to drive at 60, everyone else would still drive at 80, and so a lot of dangerous overtaking would result. For each individual the right action is to keep driving at 80, so avoid dangerously disrupting the traffic flow; yet the right group action is for everyone to drive at 60. Thus, we have in this example a right group action – everyone together driving at 60 – with each and every constituent individual action – each action of a person driving at 60 – wrong. And also we have a wrong group action – everyone together driving at 80 – with each and every constituent action – each action of a person driving at 80 – right. We see, therefore, that not even the attractive-sounding principle that if a group action is right, at least one of its constituent acts is right, is valid. (1987, 102-3)

This case presupposes some initial wrongdoing by some individuals – in particular, the initial drivers who break the speed limit – which means that the case does not show that a morally suboptimal outcome would result if *everyone* followed morality, which means that the case does not show that morality is sometimes DCSD. Furthermore, even if we imagine a group of morally flawless agents somehow 'thrown into' the case Jackson describes as in a stampede, the case still does not clearly show that morality is sometimes DCSD, and for similar reasons does not support Jackson's own conclusions.

The problem is that, contrary to what Jackson tacitly assumes, each individual driver can choose among a wide range of possible speeds. This detail undermines Jackson's argument, because although no individual driver is required to reduce his or her speed instantaneously to the morally ideal speed of 60, nonetheless at each moment each individual is required to reduce his or her speed *slightly* – which means that if everyone in the group follows morality, the morally ideal speed of 60 will be reached by the group *in the morally optimal way given the group's starting point*. (Upon reflection, it seems clear that this is what morality would require in such a case, on the assumption that it is common knowledge that morality will be universally followed.) As a result, if everyone follows morality, this leads to the morally optimal outcome of everyone driving 60, and it leads to that outcome along a path that is also morally optimal given the relevant starting point – which arguably means that if each person does follow morality along that ideal path, then the group itself also acts rightly at each moment along that path, given its suboptimal starting point. As a result, Jackson's case does not clearly support his conclusions that it is possible to

"have a group action which is wrong, yet every constituent act is right; and a group action which is right yet every constituent act is wrong". Furthermore, even if 'is wrong' is stipulated to mean 'has a suboptimal instantaneous outcome' (as Jackson intends),[23] Jackson's case is still consistent with the idea, and might even seem to illustrate the truth of the idea, that the optimal course of action for a group is in perfect harmony with the optimal course of action for each of its constituent individuals *whenever a stable equilibrium develops as a result of every individual following morality*.

It could be claimed that this *equilibrium objection* also undermines the force of the Stampede Case discussed above. However, a crucial difference is that in a stampede, in contrast to highway traffic, individuals have only two real options: continue stampeding at the dictated rate, or else be trampled – and if everyone continues stampeding at the dictated rate, then all individuals will continue to have only those two options, ensuring that the ultimate outcome never tends toward an equilibrium that is morally desirable, given realistic assumptions.[24]

More importantly, even if such an equilibrium explanation were available for the stampede cases, such an explanation is not available regarding the units of good cases discussed above, because those latter cases involve a 'one-shot' decision situation in which it is simply impossible for a desirable equilibrium to develop in the way the equilibrium objection assumes. As a result, those cases provide a decisive demonstration that morality and all other forms of normativity are sometimes dramatically DCSD, and a decisive demonstration that the best course of action for a group can radically come apart from the best course of action for each of its constituent individuals, even when a stable equilibrium develops as a result of each individual following morality.

References

Adams, R. 1997. "Should Ethics be More Impersonal?", in Dancy, J. (ed.) *Reading Parfit*. Wiley.

Gibbard, A. 1971. *Utilitarianism and Coordination*. Harvard U. Reprinted 1990, Garland.

[23] See Jackson's discussion of 'objectively right' (1987, 92).

[24] Another crucial difference is that stampedes arise without any wrongdoing by any individual, unlike Jackson's example involving high-speed highway traffic.

Jackson, F. 1987. "Group morality". In Smart, Pettit, Sylvan, and Norman (eds.), *Metaphysics and Morality: Essays in Honour of J. J. C. Smart*. New York, NY, USA: Blackwell.

Parfit, D. 1984. *Reasons and Persons*. Oxford UP.

Parfit, D. 1986. "Comments", *Ethics*.

Parfit, D. 2011. *On What Matters*, Volume One. Oxford UP.

Parfit, D. 1988. "What we together do". Unpublished m.s.

Rabinowicz, W. 1989. "Act-utilitarian prisoner's dilemmas", *Theoria*.

Rachels, S. 1989. "Counterexamples to the Transitivity of Better Than", *Australasian Journal of Philosophy.*

Regan, D. 1980. *Utilitarianism and Cooperation*. Oxford UP.

Schelling, T. 1957. "Bargaining, Communication, and Limited War", *Journal of Conflict Resolution*.

Temkin, L. 2011. *Rethinking the Good*. Oxford UP.

Krister Bykvist[1] & Karsten Klint Jensen[2]

Having It Both Ways?
On the Prospects for a Cooperation-Friendly Harmonization of Individual and Collective Maximization in Moral Hi-Lo Cases[3]

This paper analyses moral Hi-Lo Cases, which were introduced by Donald Regan's Utilitarianism and Co-operation. Moral Hi-Lo cases are moral coordination problems where coordination equilibriums are ranked by strict betterness. We argue that moral Hi-Lo cases are not just abstract hypothetical cases, there are important real-life cases of this kind, e.g., some climate change cases; and that moral Hi-Lo cases are not just a challenge for utilitarians; they are challenge for all theories that can be represented by a maximizing teleological structure. Moral Hi-Lo cases pose the challenge for individually maximizing theories that they are not collectively maximizing. We show that the widespread solution to moral Hi-Lo cases of adding the option of taking a cooperative stance to the choice situation risks changing the topic. Moreover, in the changed situation, simply making available a cooperative attitude or act is not sufficient to harmonize individual and collective maximization. This suggests that the problem sticks deeper than exclusively act-orientedness, as Regan suggested. It is not sufficient for this harmonization to assume that it is possible to influence the other agent and make her cooperative, it is necessary to actually influence her, but even with this extra assumption about actual influence, taking a cooperative stance for the best outcome may not be mandatory, if the strategy as a whole involves costs, which is a realistic assumption.

[1] Institute for Futures Studies & Stockholm University, krister.bykvist@philosophy.su.se

[2] Institute for Futures Studies, karsten.klint-jensen@iffs.se

[3] Funding from Riksbankens Jubileumsfond (grant numberP22-0662) is gratefully acknowledged.

1. Introduction

Donald Regan (1980) presented what we shall call *moral Hi-Lo problems* in his ground-breaking book *Utilitarianism and Co-Operation*. His proto-type case, which he used throughout the book, looks like this: There are only two agents in the moral universe, Whiff and Poof. Each has a button which he can push or not. The possible outcomes are evaluated by numbers representing units of value for the overall state of the world. Neither agent can influence the other's choice.

Table 1. Regan's Whiff-and-Poof-case

		Poof	
		Push	Not-push
Whiff	Push	10	0
	Not-push	0	6

Reagan was inspired by a similar case, set up by Allan F. Gibbard (1965). As Gibbard saw things, such coordination problems pose a challenge to act utilitarianism, since it does not necessarily ensure the collectively best outcome. If Poof not-pushes, act utilitarianism requires Whiff to not-push as well. And the same holds for Whiff if Poof not-pushes. In other words, the act pattern (not-push, not-push) is individually maximizing, i.e., the best each agent could do on their own. But they could together bring about an outcome of value 10 by each pushing. Thus, the act pattern (push, push) is collectively maximizing, i.e., the best they could do together. The example thus shows that an individually maximizing act-pattern need not be collectively maximizing. Gibbard concludes that some form of institutional coordination is needed to achieve the collectively best outcome. However, Regan's aim is to demonstrate that coordination *can* be achieved by morally motivated agents.

The climax of Regan's analysis is the proof that no theory which fulfils a necessary condition for being exclusively act-oriented can be strongly collectively maximizing.[4] The necessary condition for a theory to be exclusively act-oriented in a Hi-Lo case is that the theory specifies, for each agent, some subset of the set of available acts, such that the agent satisfies the theory iff she does an act from the specified subset.

This led Regan himself to suggest that act utilitarianism should be supplemented by a somewhat complicated decision procedure, which he argues is able to ensure coordination for the collectively best outcome. No one else has followed him in that. But many have accepted the premise that a solution must involve going beyond an exclusively act-oriented theory, e.g. by adding a cooperative attitude or an act inviting

[4] We define the properties of individually and (strongly) collectively maximizing theories below.

to cooperation to the case and supplementing an act-consequentialist type of theory with a duty to take on the attitude and/or perform the invitation.

One of the aims of this paper is to make a rational reconstruction of this type of theory in order to explore the prospects of finding a cooperation-friendly harmonization of individual and collective maximization in moral Hi-Lo cases. More specifically, we shall show the following:

1. moral Hi-Lo cases are not just abstract hypothetical cases, there are important real-life cases of this kind, e.g., some climate change cases (sections 2 and 3);
2. moral Hi-Lo cases are not just a challenge for utilitarians; they are a challenge for all theories that can be represented by a teleological structure (sections 4 and 5);
3. adding the option of taking a cooperative stance to the Hi-Lo case in Table 1 risks changing the topic. Hence a solution to the changed Hi-Lo case may not be a solution to the original case (section 5);
4. in the changed case, simply making available a cooperative attitude or act is not sufficient to harmonize individual and collective maximization. This clearly suggests that the problem sticks deeper than exclusively act-orientedness, as Regan suggested (section 6);
5. in the changed case, it is not sufficient for this harmonization to assume that it is *possible* to influence the other agent and make her cooperative, it is necessary to *actually* influence her (section 6);
6. but even with this extra assumption about actual influence, taking a cooperative stance for the best outcome may not be mandatory, if the strategy as a whole involves costs, which is a realistic assumption (section 6).

Before we start arguing for these claims, we shall first give a more precise definition of a moral Hi-Lo case.

2. What is a moral Hi-Lo problem?

Moral Hi-Lo problems constitute a subclass of what can be called moral coordination problems. Regan does not provide a general definition of either of these; he mainly works from the generic case cited above. Let us first adapt from game theory the concept of a coordination equilibrium[5] to this context:

[5] Lewis (1969: 14). The concept is clearly modeled on the concept of a Nash equilibrium, which is defined in the

> A *coordination equilibrium* is a combination of acts in which the overall state of the world would not be better if any one agent alone acted otherwise.

We shall define a moral coordination problem thus:

> In a *moral coordination problem*, each of *n* agents chooses one act from a finite set of alternatives. Each outcome has an objective moral value. There are at least two coordination equilibria. Coordination equilibria are ranked by an 'at least as good as'-relation, and no non-equilibrium combinations is better than any equilibrium.

We get a *moral Hi Lo problem*[6] if none of the equilibriums are equally good; i.e. all equilibriums are ranked by strict betterness. Regan only considers simple two person cases with two available acts and we shall follow him in that. This simplifies the discussion considerably and can for the most part be done without any loss of generality.

It might be relevant to add to the ranking of outcomes an assessment of the value difference between the best equilibrium outcome and the second-best, and between the second-best equilibrium outcome and best non-cooperative outcome. Let us call it a *high-stake case*, when either or both of these differences are significant.

Regan's Whiff-Poof case is presented *as if* it were a game, more precisely a coordination problem. And clearly, the case shares with games the property that the outcome of an agent's choice is depending on the choices made by others. But the case cannot simply be identified with a standard game. The numbers represent an agent-neutral ranking of outcomes as overall states of the world. They do not necessarily represent the preferences of the agents, as they would do, if it was a game.

As Regan presents the case, there is uncertainty about the preferences of the agents, whereas in a standard game, full information about preferences is assumed. The agents in the situation may have both self-interested preferences and conflicting preferences, resulting in rankings of the outcomes which deviate from the objective moral ranking assumed in the case.

It is part of Regan's proposed decision procedure to provide shared information about the agents' preferences. Thereby the decision problem, initially under uncertainty, can be transformed into a moral coordination problem, which however only may be faced by the subgroup who shares an agent-neutral objective moral ranking and disregards those who are unwilling to cooperate.

Moral Hi-Lo cases show up in a more *indirect* way as well, for some standard games can be transformed to a moral Hi-Lo case. Consider a standard Hi-Lo game (the first

context of non-cooperative games, e.g. Luce & Raiffa (1957: 106).

[6] The name is of course inspired by Bacharach (2006).

number represents the preference intensities of agent A, the second those of B) (Bacharach 2006):[7]

Table 2. A standard Hi-Lo case

		B	
		Hi	Lo
A	Hi	5 / 5	0 / 0
	Lo	0 / 0	3 / 3

Bacharach calls this a *common interest* game. If (Hi, Hi) is better than (Lo, Lo), which is better than (Hi, Lo), we have a moral Hi-Lo problem. This would be so, if the preference intensities were simply summed.

It is important to note, however, that Hi-Lo cases are relevant to *non-utilitarian* moralities as well. The numbers need not represent the sum total of preference intensities or wellbeing; they can represent some other aggregation of all relevant values (not necessarily welfarist). Nor do we have to assume consequentialism for Hi-Lo cases to be of interest. As Portmore (2018) points out, the numbers can be seen as representing the moral value of the act-combination and its associated outcome. Even deontologists and virtue-ethicists can accept that things have impartial moral value. But, pace Zimmerman (1996) and Pinkert (2015), impartiality or agent-neutrality is not required either. Here is a moral Hi-Lo case where the numbers represent agent-relative moral value, degrees of moral reasons or choice-worthiness, which can depend on the agent's motivations or dispositions. The first number represents the agent-relative value for A of A's action (and its outcome, if that matters), the second number the agent-relative value for B of B's action (and its outcome, if that matters).

Table 3. A moral Hi-Lo case with agent-relative values

		B	
		Hi	Lo
A	Hi	-2 / 5	-8 / 0
	Lo	-8 / 0	-4 / 2

In this case, the available actions do not even have the same polarity for the agents. All actions are bad relative to A, and neutral or good relative to B. Still, if (Hi, Hi) is

[7] Throughout, when we for simplicity call acts 'Hi' and 'Lo', they need not represent the same act for the two agents. 'Hi' for each represents the act which combines to the best equilibrium, and 'Lo' for each agent represents which combines to the second-best equilibrium.

better than (Lo, Lo), which is better than both (Hi, Lo) and (Lo, Hi), we have a moral Hi-Lo problem.

Bacharach (2006) shows that some games of conflict can be transformed into common interest Hi-Lo games, which by the argument above then can be transformed into a moral Hi-Lo case. Stag Hunt is one example:

Table 4. Stag Hunt

		B	
		Stag	Rabbit
A	Stag	2 / 2	-1 / 1
	Rabbit	1 / -1	1 / 1

This becomes a moral Hi-Lo problem, if (Stag, Stag) is better than (Rabbit, Rabbit), which is better than both (Rabbit, Stag) and (Stag, Rabbit). This would be the case if the value is the sum of individual preference intensities

Other examples are some versions of the Prisoner's Dilemma, like the following one.

Table 5. Prisoner's Dilemma

		B	
		Cooperate	Defect
A	Cooperate	4 / 4	0 / 5
	Defect	5 / 0	3 / 3

This will become a moral Hi-Lo problem, if (Cooperate, Cooperate) is better than (Defect, Defect), which is better than both (Cooperate, Defect) and (Defect, Cooperate). This would be the case if value is the sum of preference intensities.

3. Why are moral Hi-Lo problems important?

That many important decisions involve coordination problems is widely accepted. Choices of great importance often require coordination to achieve the optimal results. Some of these problems can be seen as moral high stake Hi-Lo cases, either involving several agents or only two. There are numerous examples of such Hi-Lo cases, both from the individual and the political spheres. For example, many rescue cases have this form. They can be found whenever there are two rescue options, both of which requires cooperation to succeed, and where one is better than the other (cf. Colman et al. 2014: 36). Suppose that in the aftermath of an earthquake we find out that there is one person buried under a crumbled building and several people buried under

another. There is no time to rescue all people, because the buildings are far apart and the oxygen levels under the buildings are falling quickly. In order to successfully save any of the people, we need to both be there to move the heavy rubble. Our radios are not working so we cannot communicate. This case has the form of a moral high stake Hi-Lo case. We can either together save the larger group (Hi, Hi) or the lone person (Lo, Lo), but no one will be saved if we go to different buildings, (Hi, Lo) or (Lo, Hi).

Other examples come from the climate change context. The most straightforward and common cases are choices between climate actions that require extensive infrastructure to work and where each party's contribution to the infrastructure is crucial. One pair of coordinated climate options (Hi, Hi) might be better than another (Lo, Lo), but (Hi, Lo) and (Lo, Hi) would be worst, because unilaterally going for one option would provide an insufficient infrastructure. To take a mitigation case, 'Hi' can be electrification of aviation and 'Lo' expansion of fast trains. Both parties going for electrification of aviation is better than both going for expansion of fast trains, because of the time benefits of flight travel. But unilateral choice would not provide sufficient infrastructure for either aviation or expansion of fast trains. Instead, the unilateral choice would only incur futile costs. Similar cases can be constructed in which the choice is between different energy systems, for example, hydro and wind versus nuclear.

Another case is an adaption case, where the choice is between two adaption strategies against flooding: building seawalls and relocating the population. Both parties going for building seawalls is better than both going for relocation, because no one is forced to move if seawalls are built. However, the mixed options are worse because then not enough infrastructure will be put in for the seawalls to be effective and people will still have to relocate. To make things more concrete, assume that two neighbouring nations are each threatened by sea-level rise, where they share a salient geographic border. Without both nations building seawalls there would be flooding. If only one nation builds a seawall, the flood waters will just be pushed toward the part of their shared geographic boundary that is uncovered, and the same damage will occur, with the water just taking a different (slower) route. The country that built the wall will eat the entire cost of building the wall, which will then be useless.

As pointed out above, moral Hi-Lo cases may show up in a more indirect way as well, when a decision we face have the form of a Prisoner's Dilemma that can be transformed to a moral Hi-Lo case. One famous illustration is the so-called Polluter's Dilemma, which has the structure of a Prisoner's Dilemma.[8] Suppose A and B (who can be nations or individuals) have two options: not pollute (cooperate) and pollute

[8] For a thorough discussion of when a pollution case is best seen as a case of a Prisoner's Dilemma, see Pellikaan and van der Veen (2002).

(defect). The Pareto-optimal outcome is one in which both do not pollute. But it is not an equilibrium, since each agent would be better off unilaterally defecting and polluting. This assumes that the pollution produced by each agent is not significant enough to outweigh the benefits of not having to pay for making production pollution-free.[9] To get a moral Hi-Lo case it is enough to assume that both cooperating and not polluting is better than both polluting, which is better than one polluting and the other not polluting. In some cases, e.g. climate cases, the agents in the decision problem may not stand to benefit themselves but all benefits go to a third party (future generations). The problem would then not have the structure of a traditional Prisoner's Dilemma, but it could still be transformed into a moral Hi-Lo problem, if the agent-neutral ranking is this: (Hi, Hi) is better than (Lo, Lo), which is better than both (Hi, Lo), (Lo, Hi).

4. Assumptions and definitions

In the following we shall clarify the framework for our discussion, which we largely adapt from Regan (1980) who has framed the subsequent discussion. We are dealing with *objective moral theories*, according to which rightness depends on the facts, not the agent's beliefs or evidence about the facts. In contrast to game theory and decision theory, information and subjective probabilities are not normatively relevant.[10]

It is true that many authors find it necessary to add more information to the agents about the cases. As pointed out above, Regan's own strategy is to transform the moral Hi-Lo case into a coordination game where the agents have information about its structure. However, since our analysis is focusing on objective duties, we do not need such assumptions.

Second, we start from an individual duty perspective. The question is what each individual ought to do. As we shall see, some authors want to introduce collective duties, and our starting point does not exclude this possibility, since one could argue that individual duties can be derived from collective ones.

Third, in contrast to Regan who is concerned with act utilitarianism only, we are dealing more broadly with moral theories that are teleological in the following weak sense: theories that can be given a maximizing teleological representation, where rightness of an action is determined by the highest-ranked outcome (at least if other

[9] It is controversial whether this applies to national agents in climate change, since one could argue that here the pollutions are significant, at least for big emitters such as the US and China. For overviews of the relevance of Prisoner's Dilemma to climate change, see Chander (2018) and Magli and Manfredi (2022).

[10] Regan allows for objective probabilities as well, but as pointed out by Rabinowicz (1989), this creates more problems than benefits. Since probabilities play no role in his argument anyway, we ignore this possibility.

things are equal). This outcome can be an outcome of a group-action of which the individual action is a part.

Finally, we assume that each available alternative for an individual agent has a determinate morally relevant outcome given the pattern of behavior of the other agents.

We shall also list some properties of moral theories (denoted T) relevant for our arguments.

- T is *universally satisfied* in some pattern of actions iff all agents do what T requires of them in this pattern.
- T is *individually maximizing* (IM) iff for any agent, in any choice situation, if the agent satisfies T in that situation, he produces by his act the 'best' consequences he can possibly produce in that situation. This is a generalization of Regan's PropAU.
- T is *collectively maximizing* (CM) iff for any pattern of actions in which T is universally satisfied, the class of all agents produce by their acts taken together the 'best' consequences that they can possibly produce by any pattern of behavior. This is Regan's PropCOP, which is one version of what is often called moral harmony.
- T is *strongly collectively maximizing* (SCM) iff for any situation involving choices by any number of agents, the agents who satisfy T in that situation produce by their acts taken together the 'best' consequences that they can possibly produce by any pattern of behavior, given the behavior of agents who do not satisfy T. This is Regan's property of T being Adaptable. As pointed out by Regan (1980: 107), 'T is SCC' entails 'T is CM'.

The reason we talk about best in square quotes is that it is meant to also capture the case where the relevant consequences are *at least as good* as that of any other relevant alternative.

Note that the property of being IM is defined in an unqualified way, i.e., for *any* choice situation. We shall later discuss the prospect of theories being IM only in certain specific choice situations.

5. Why are moral Hi-Lo cases a problem for act utilitarianism? Generalizing the challenge and two strategies to address it

Regan presented his Whiff/Poof case as a problem for act utilitarianism. We have already hinted at what the problem is: In the Whiff/Poof case, given that the agents act independently of each other, there are two patterns of actions, where act utilitarianism is universally satisfied, (Push, Push) and (Not-push, Not-push). In (Not-push, Not-push), the agents together do not produce the best possible consequences they possibly can. Hence, act utilitarianism is not CM.

However, there is a simple generalization of the challenge at hand. Consider any theory which is individually maximizing applied to a two-person moral Hi-Lo problem. Any such theory is universally satisfied in the pattern (Lo, Lo), and therefore it is not collectively maximizing. Hence, no theory which is individually maximizing can be collectively maximizing. This result should not be surprising, since (Lo, Lo) is a coordination equilibrium and the concept of a coordination equilibrium is defined from an individually maximizing perspective.

There seem to be two possible strategies to address this challenge. One is to reach for a theory which is only CM and thus giving up the requirement that it should also be IM. Call this the *non-reframing approach*. For this approach, the standard two-person choice situation still has only two options, Hi or Lo. We could imagine both individualist and collectivist versions of this strategy. As an individualist example, consider a simple rule consequentialism, according to which you should always follow the best rule, i.e. the rule that would have the best consequences if everyone followed it. This theory would tell each agent to do Hi, since the best rule would require both to do Hi. But consider also a simple collectivist version, which would tell each person to do their part in the collective duty to do the best we can do together. Since the best we can do together is that both do Hi, each agent ought to do Hi.

The non-reframing approach is really a non-starter, however, since it would implausibly require each agent to do Hi, even if a *catastrophe* would ensue. In fact, we are uncertain if anyone seriously would want to defend it. There is one qualification, though, because the argument is based on high-stake situations. Imagine a low-stake situation, where the value of the third-best outcome is not very much worse than the value of the best. Here, there could be a hard choice whether or not we should accept that a collectively maximizing theory might involve some relatively minor costs or harms for individuals. It is defensible to accept such costs in low-stake situations. Nonetheless, a collectively maximizing theory can only be plausible if it rejects such

costs in high-stake situations, which is to accept that collective maximization does not apply universally.

Hence, the conclusion is that no theory should be CM. But given low-stake cases, perhaps no theory should be IM either. The best theory might combine individual and collective elements. It is clear, however, that a theory should agree with individual maximization in high-stake cases.

The other strategy we call the *reframing approach*. This approach accepts that IM theories cannot be CM in the standard moral Hi-Lo case. The aim is to reframe or transform the standard case to a situation where an IM theory *can* be CM. This reframing is done by altering the existing options of the original case or adding one or more options to them and thereby creating new compound alternatives. As we will see, these new compound alternatives need not be compound *actions* (strictly speaking); there may instead be combinations of attitudes and actions, or reasoning processes and actions. Though this approach confirms that the original challenge cannot be met, it might of course still be of interest to see if a theory can be both IM and CM in such reframed and perhaps more realistic Hi-Lo cases.

There are many different instances of the reframing approach, including the ones defended by Regan (1980), Zimmerman (1996), Portmore (2018), Pinkert (2015), Schwenkenbecher (2021), and Goodin (2012). Let us start with quoting a short summary of Regan's (1980: 135f.) approach, which appears to have been an inspiration for most re-framers (italics added):

"Each agent should

- *be willing* to take part in a joint attempt to produce the best consequences possible by co-ordinating his behaviour with the behaviour of other agents who are willing

- *consider* the other agents involved in the co-ordination problem he is making a decision about and determine which of those other agents are available to be co-operated with.

- *ascertain* how other agents who are not (for whatever reason) available to be co-operated with are behaving or disposed to behave

- *identify* the best possible pattern of behaviour for the group of co-operators […] given the behaviour (or dispositions to behave) of the non-co-operators

- *do* his part in the best pattern of behaviour just identified."

Regan presents his theory as a non-exclusively act-oriented IM theory which is also generally SCM.[11] It is designed as a decision procedure which determines which of the acts available in the original situation (Hi or Lo) is the right one to choose under the circumstances. But it is clear that he is re-framing the situation, since he adds that each agent should not just perform a certain action but also have a certain willingness, and moreover consider, ascertain, and identify aspects of the aspects of the choice situation.

The original options, i.e. simply choosing without going through (or fail or reject) the procedure, are not mentioned by Regan and presumably they are not available. Instead, the theory tells you to go through a procedure, consisting of a number of preparatory steps and concluding in the actual choice of an action.

Zimmermann is also re-framing. He presents a revised version of his preferred individually maximizing theory, according to which "(d)oing the best one can must be accompanied by the adoption of a certain *attitude*" (italics added).[12] This is a clear alteration of the options in the original Hi-Lo case.

Here are other examples of what re-framers propose:

- Be willing to cooperate with others who are willing (Sugden 2015)
- Form a disposition to cooperate (Portmore 2018; Pinkert 2015)
- Make a cooperative commitment (Goodin 2012)
- Take a cooperative attitude (Zimmerman 1996)
- Do we-reasoning (Bacharach 2006; Schwenkenbecher 2021)
- Say or signal that one is cooperative (Schwenkenbecher 2021)
- Identify cooperators (Regan 1980)

We suggest that these proposals can be summarized as adding a new option of 'taking a cooperative stance' (Cop). (Note that this extra option need not be an action, strictly speaking.) So, instead of just having the alternatives doing Hi or doing Lo, we now have the alternatives:

[11] Even though the procedure involves acquiring certain beliefs, Regan boldly claims that it is a completely objective theory. But this appears doubtful, since acting rightly according to the theory depends on having the correct beliefs.

[12] Zimmerman (1996: 263).

- doing Hi/Lo while or after[13] taking a cooperative stance (Cop&Hi, Cop&Lo)
- doing Hi/Lo while or after refusing to take a cooperative stance (¬Cop&Hi, ¬Cop&Lo)

We assume that taking a cooperative stance entails that one is *successfully cooperative* (in the pursuit of value). Of course, this is not always true. One can take a cooperative stance and fail to be cooperative, because of weakness of will, dishonesty, or lack of crucial information. But we want to make the best case for the re-framers. We will assume that in the Hi-Lo two-agent case, being successfully cooperative entails being such that

- one would do Hi, if the other agent were to do Hi.

Note that to make taking a cooperative stance available in the choice situation is to reject the assumption which is a defining characteristic of Regan's original problem, namely that the agents are *disconnected* in the sense that they choose independently of each other. For if one agent takes a cooperative stance, this is one way she can connect to the other agent in the sense that if the other agent did Hi, she would respond with Hi.[14]

Refusing to connect is being such that one chooses independently of the other agent. We shall in particular be concerned with the case where both agents do Lo, in which case not taking a cooperative stance entails that:

- one would do Lo, even if the other agent were to do Hi.

In the general case of any group of agents G, a member of G taking a cooperative stance towards G entails

- one would do Hi, if all other members of G were to do Hi.

For our argument we only need these plausible ways of (not) taking a cooperative stance. Furthermore, we shall conduct our argument concentrating on the standard two-person case. In this case, the reframed situation looks like this:

[13] For some re-framers, it is important that taking a cooperative stance (or not) may be at a time earlier than the choice between 'Hi' and 'Lo' so the other may be able to respond to the stance. We discuss cases where this issue becomes important below.

[14] It is also possible to attempt to connect by taking an *uncooperative* stance, i.e. fulfilling the condition 'One would do Lo if the other agent were to do Hi, and one would do Hi, if the other agent were to do Lo'. Bykvist (forthcoming) calls this 'the contrarian option'. It was first envisaged by Feldman (1986). It plays no role in our argument, and we shall therefore not include it in the model.

Table 6. The re-framed choice situation

			B			
			Cop		¬Cop	
			Hi	Lo	Hi	Lo
A	Cop	Hi	10	n.a.	10	0
		Lo	n.a.	6	0	6
	¬Cop	Hi	10	0	10	0
		Lo	0	6	0	6

N.a. (not applicable) signifies an impossible combination. A cannot fulfill the Cop-condition 'if B were to do Hi, then A would do Hi', if A does Cop&Lo when B does Cop&Hi, and vice versa.

This matrix makes it clear that re-framers are changing the topic, simply because the availability of new alternatives reflects that the original assumption of independent choices has been rejected. In other words, the reframed situation is a changed situation. Could it nevertheless be argued that these alternatives were available in the original situation, in other words, that the characterization of the original case ignored certain available options?

It is clear that the original case has been changed, if taking a cooperative stance is identified with a *physical* action, like saying or signaling that one is cooperative, identifying cooperators, or making others cooperative. These physical actions need not be available to an agent, not even implicitly. But perhaps the option of taking a cooperative stance can be seen as implicit in the original case, if they are identified with *mental* acts, such as doing we-reasoning. This is not clear, however, since on the most famous account of we-reasoning (Bacharach's 2006), such reasoning is not a voluntary option.

There is no need for us to decide exactly when the original case can be said to be changed, since our main aim is to examine under which conditions satisfaction of a theory being IM is compatible with it being CM.

6. The prospects for a cooperation-friendly harmonization of individual and collective maximization

6.1 The Nice Case

What Regan thought he could achieve with his decision procedure is a coordination game where both take a cooperative stance. This is true. See Table 7.

Table 7. The nice case

		B	
		Cop&Hi	Cop&Lo
A	Cop&Hi	10	n.a.
	Cop&Lo	n.a.	6

In this situation, if B does Cop&Hi, it is IM for A also to do Cop&Hi. In fact, there is no other choice available! The same holds for B. Any IM theory is universally satisfied in the pattern (Cop&Hi, Cop&Hi) and thus trivially satisfies being CM. But note this only holds, if taking a cooperative stance involves no costs, a topic we will come back to below.

6.2 The challenge from disconnection

Suppose that there is a mutual disconnection between the agents in the sense that choosing Cop would not make the other agent choose Cop. Suppose further that both A and B choose ¬Cop&Lo (marked in bold in table 8). Consider Table 8. In the yellow area, you find the nice case. The grey column shows the available outcomes from A's point of view, when B does ¬Cop&Lo, and A cannot do anything to make B choose Cop. As can be seen, even if A were to do Cop, the best she can do in that case is also to choose Lo (the value would be 6, which is marked by green), since her cooperative stance would have no effect on B. In other words, A can do nothing to move B into the nice case. Choosing ¬Cop&Lo would have the same value (again with value 6, marked green). The same holds for B, since the situation is exactly symmetrical. Hence, we get the result that any theory that is IM in this situation is universally satisfied in the pattern (¬Cop&Lo, ¬Cop&Lo). But this pattern is not CM. We are in effect back to the predicament of the original case where both agents do Lo and no agent can influence the other agent's actions. Furthermore, note that there is a tie for each agent between choosing (¬Cop&Lo) and (Cop&Lo) when it comes to IM, since each option would lead to an outcome with value 6.

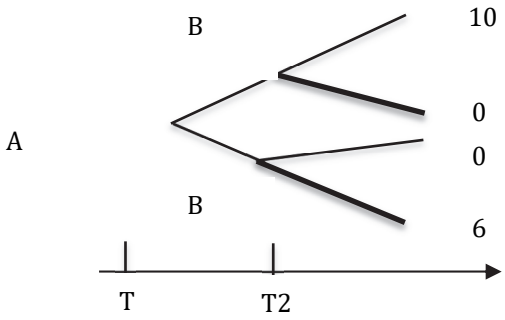
Table 8. The case of disconnection

			B			
			Cop		**¬Cop**	
			Hi	Lo	Hi	Lo
A	Cop	Hi	10	n.a.	10	0
		Lo	n.a.	6	0	6
	¬Cop	Hi	10	0	10	0
		Lo	0	6	0	6

Would things change if *one* of the agents took a cooperative stance? Many re-framers think that in that case the non-cooperative agent would no longer be an individual maximizer, since the cooperative agent is such that she would respond with a Hi to a Hi. The cooperative agent should then choose what is IM under the circumstances, which in the case of Table 8 would be Cop&Lo.[15]

But it need not be true that the non-cooperative agent is not IM if we consider a *diachronic* case in which the cooperative agent acts first. In this case, it would be IM for the cooperative agent to choose Cop&Lo, since the other agent is uncooperative and would choose Lo, no matter what she did. Now, once Lo has been chosen by the cooperative agent, the uncooperative agent faces a choice between Hi (with or without Cop), which would lead to an outcome of value 0, and Lo (with or without Cop), which would lead to an outcome of value 6. So, choosing ¬Cop&Lo would be an IM option for the uncooperative agent. So, if the cooperative agent chooses Cop&Lo and the uncooperative chooses ¬Cop&Lo, they are each choosing what is IM. But this act-pattern is not CM. The situation can be illustrated with the following tree diagram:

Figure 1. A diachronic case



Suppose A is the cooperative agent, and B is the uncooperative agent. A bold line indicates what is actually chosen or would be chosen. Going up for A means choosing Cop&Hi; going down for A means choosing Cop&Lo. Going up for B means choosing Hi (with or without Cop); going down for B means choosing Lo (with or without Cop). (We have omitted the branches that involve A choosing ¬Cop, since they do not make a difference for the argument.) It is clear that if A chooses Cop&Lo at T1, this is an individually maximizing choice. Once this has been done, B would

[15] In a larger group, an individually maximizing theory would still be SCM when the subgroup of cooperative agents together chooses the best outcome, given the behavior of the non-cooperative agents. Regan (1980) considers this feature a major advantage of his theory.

be an individual maximizer if she were to choose ¬Cop&Lo at T2. But this act combination is not CM.

Note that how this situation differs from the *synchronic* one, where both agents act at the same time. In that case, since A is cooperative and would do Hi if B did Hi, B does not perform an IM action by doing Lo. If she had done Hi instead, A would have done Hi and the outcome of value 10 would have been achieved. This also holds for the diachronic case where the uncooperative B acts first. If B chooses Lo at T1, she does not act in an IM manner, since if she had done Hi instead at this time, A would have followed it up later at T2 with a Hi and the outcome with value 10 would have been achieved.

6.3 Adding the value of being cooperative

Many re-framers work from the intuition that refusing to cooperate (choosing ¬Cop) is wrong. They are willing to accept that if they face someone who would not cooperate and chooses Lo, the best you can do is to choose Lo. But they find it hard to "allow two wrongs to make a right" (Zimmermann 1996: 257), which seems to be the case for a theory which is universally satisfied by the pattern (¬Cop&Lo, ¬Cop&Lo). Hence, in that case, the theory should require Cop. It does not make the outcome better,[16] but it allows the theory not to violate the property of being SCM.

But note that a theory which merely stipulates that refusing to cooperate (choosing ¬Cop) is wrong jeopardizes the property of being IM. One might think that one could address this problem by assigning some extra final value v to an outcome if it results from Cop rather than ¬Cop. The matrix would then look like this.

Table 9. Added value of being cooperative

			B			
			Cop		¬Cop	
			Hi	Lo	Hi	Lo
A	Cop	Hi	10+2v	n.a.	10+v	0
		Lo	n.a.	6+2v	v	6+v
	¬Cop	Hi	10+v	v	10	0
		Lo	v	6+v	0	6

Note that by adding this value v we are breaking the tie for A: ¬Cop&Lo is no longer an IM choice; Cop&Lo would have a better outcome with value 6+v. Suppose that B is choosing ¬Cop&Lo and would do so, no matter what A did. Then choosing Cop&Lo is the only IM choice for A, as can be read from the grey column. But this

[16] As pointed out by Feldman (1986).

maneuver does not help in general. If A acts first, then B will later have a choice between doing Cop&Lo with value $6+2v$, doing ¬Cop&Hi with value v, and doing ¬Cop&Lo with value $6+v$. So, choosing Cop&Lo would be an IM choice for B. But the combination (Cop&Lo, Cop&Lo) is not CM. So, even if we have managed to break the tie for A, we still have not succeeded in establishing harmony between IM and CM in this case.

6.4 A stronger form of connection

We have seen that we cannot combine CM and IM in all cases where there is *one* agent who is not taking a cooperative stance. Let us now look at cases where one can do Cop and at the same time make sure that the other agents take a cooperative stance, perhaps by 'getting assurance' that the other person will cooperative, as in Sugden (2017), 'promote' cooperation as in Cripps (2013), or create a collective agent as in Collins (2019). Consider the situation in Table 10 from A's point of view, where 'Cop*' is defined as

- one would do Cop *and* at the same time make sure the other agent does Cop.

Table 10. Making the other agent taking a cooperative stance

			B					
			Cop*		Cop		¬Cop	
			Hi	Lo	Hi	Lo	Hi	Lo
A	Cop*	Hi	10	n.a.	10	n.a.	n.a.	n.a.
		Lo	n.a.	6	n.a.	6	n.a.	n.a.
	Cop	Hi	10	n.a.	10	n.a.	10	0
		Lo	n.a.	6	n.a.	6	0	6
	¬Cop	Hi	n.a.	n.a.	10	0	10	0
		Lo	n.a.	n.a.	0	6	0	6

Suppose B does not take a cooperative stance, but chooses ¬Cop&Lo. If A were to take a strong cooperative stance (Cop*) towards B, she would make B take a cooperative stance towards her and thereby move her into upper right orange area doing Cop&Hi. The outcome would then be 10 (blue). Any other act by A would have a worse outcome (red). (It is possible that B would even want to do Cop*&Hi, thereby moving into the green area).

Suppose the same holds from B's point of view. Then any IM theory is universally satisfied in the patterns (Cop*&Hi, Cop&Hi), (Cop&Hi, Cop*&Hi) and (Cop*&Hi, Cop*&Hi), and thus also CM. Hence, we have found other cases than the 'nice case' (yellow area) where there is harmony between IM and CM. But it is important to note

that merely adding the option of taking a cooperative stance (Cop) does not help. Nor does it help to assume that one agent is taking the cooperative stance, as we pointed out in section 6.2. We also need to assume that the agents' Cop-actions are *mutually* connected.

This example can also be used to fulfil another aim of this paper: to show that cooperation is not necessarily involved in the act-patterns that are both IM and CM. For suppose both A and B choose ¬Cop&Hi. Any individually and collectively maximizing theory is universally satisfied in this pattern too (green). In order to make (Hi, Hi) combinations involving one or two Cop*-stances better solutions, these outcomes need to be assigned an extra final value v, perhaps because being cooperative itself has final value, as we discussed in section 6.3.

6.5 Taking a costly cooperative stance

So far, we have assumed that taking a cooperative stance is cost-free. But this is a very unrealistic assumption. There are two ways, in which taking a cooperative stance might involve costs to the outcome. One is that it may have unintended negative moral consequences, which must be subtracted from the value of the outcome.[17] This is something almost any kind of act risks, with some probability. We shall ignore this possibility in our argument, since it is not relevant in our framework which assumes that each available alternative for an individual agent has a determinate morally relevant outcome given the pattern of behavior of the other agents.

We find it more important that taking a cooperative stance, in most cases at least, appears *certain* to involve a personal cost in terms of spent time and energy for the agent who undertakes it.[18] If only this cost is positive, however small, it will have a dramatic effect on the evaluation of the outcomes in the cases where successful coordination is achieved. Consider Table 11:

[17] Regan himself (1980: 267ff.) introduced the possibility of a mad telepath, but he did not consider it a serious problem. However, most commentators have used this example against him.

[18] Regan (1980: 267ff.) also considers costs of this kind, but – as Zimmerman (1996: 260) says – he *deliberately* ignores them.

Table 11. The cooperative stance involves a cost

			B			
			Cop(*)		¬Cop	
			Hi	Lo	Hi	Lo
A	Cop(*)	Hi	10-2*c*	n.a.	10-*c*	-*c*
		Lo	n.a.	6-2*c*	-*c*	6-*c*
	¬Cop	Hi	10-*c*	-*c*	10	0
		Lo	-*c*	6-*c*	0	6

Remember there are two ways to achieve coordination: either if both agents take a cooperative stance (what we called the 'nice case' above) or if one agent takes the strong form of a cooperative stance (cop*). Now we assume that taking a cooperative stance (choosing Cop(*), i.e., either Cop or Cop*) involves a (morally relevant) personal cost of *c*. We contrast choosing Cop(*) with choosing ¬Cop, which serves as the baseline in terms of costs.

It is clear from Table 11 that no IM theory is universally satisfied in the pattern (Cop&Hi, Cop&Hi). It is also not the pattern where the agents together produce the best possible consequences they possibly can; this price goes to the pattern (¬Cop&Hi, ¬Cop&Hi) (blue). Ironically, a theory which were to be CM in this situation would have to recommend the agents *not* to take a cooperative stance.

Finally consider how a costly cooperative stance affects 'breaking the tie'. Remember that the tie came up if one agent were to choose ¬Cop. Then the other could equally well choose ¬Cop as Cop. This tie could be broken by assigning some final value to choosing Cop. Then (see *Table 12*), if one agent were to choose ¬Cop&Lo, the IM choice by the other would be Cop&Lo, and similarly, if one agent were to choose ¬Cop&Hi, the IM choice by the other would be Cop&Hi.

Table 12. Breaking the tie

			B			
			Cop		¬Cop	
			Hi	Lo	Hi	Lo
A	Cop	Hi	10+2*v*	n.a.	10+*v*	*v*
		Lo	n.a.	6+2*v*	*v*	6+*v*
	¬Cop	Hi	10+*v*	*v*	10	0
		Lo	*v*	6+*v*	0	6

But now assume that Cop involves a (morally relevant) personal cost *c*. This means that every *v* should be replaced with *v-c* (Table 13). There are three cases to consider.

Suppose first that $v>c$. Then the situation in *Table 12* is not affected, the tie is still being broken (green combinations in Table 13).

Table 13. Breaking the tie in spite of costs

			B			
			Cop		¬Cop	
			Hi	Lo	Hi	Lo
A	Cop	Hi	10+2*v*-2*c*	n.a.	10+*v*-*c*	*v*-*c*
		Lo	n.a.	6+2*v*-2*c*	*v*-*c*	6+*v*-*c*
	¬Cop	Hi	10+*v*-*c*	*v*-*c*	10	0
		Lo	*v*-*c*	6+*v*-*c*	0	6

Suppose next that $v=c$. Then the tie is back (Table 9 above), because v and c cancel each other out.

Finally, suppose that $c>v$. The colors in Table 14 are changed from Table 13 to match this situation.

Table 14. Attempting to break the tie involves a net cost

			B			
			Cop		¬Cop	
			Hi	Lo	Hi	Lo
A	Cop	Hi	10+2*v*-2*c*	n.a.	10+*v*-*c*	*v*-*c*
		Lo	n.a.	6+2*v*-2*c*	*v*-*c*	6+*v*-*c*
	¬Cop	Hi	10+*v*-*c*	*v*-*c*	10	0
		Lo	*v*-*c*	6+*v*-*c*	0	6

Suppose the other agent were to choose ¬Cop&Lo. Then the IM choice by the other would be to likewise choose ¬Cop&Lo. Any IM theory is universally satisfied in this combination; but it is not CM. Hence, we are back in the original problem. A costly Cop does nothing to overcome it. On the contrary, in (¬Cop&Hi, ¬Cop&Hi) any IM theory is universally satisfied, and this combination is also CM. Any unilateral choice from a cooperative stance will make things worse.

7. Concluding Remarks

The analysis shows that the challenge raised by the original moral Hi-Lo problems stems from the fact that the agents choose independently of each other and are unable to influence each other. In this case, any IM theory is universally satisfied if each agent chooses Lo, but no such theory can be CM.

An attempt to give up the requirement that a theory should be IM and go for CM theory instead would lead to a catastrophe in high-stake Hi-Lo cases, and is thus not a viable option. However, it may be an option in low-stake cases, if the costs are not considered prohibitive.

It seems a better prospect to break the independence by attempting to connect cooperatively with the other agent in some way. However, this amounts to reframing of the choice situation and is thus changing the topic, at least if the reframing involves adding physical actions.

In the 'nice case', where each agent takes the cooperative stance, the IM option is also CM. For this reason, it seems natural to suggest that there should be a duty to take a cooperative stance. But if taking a cooperative stance involves a cost, the pattern where both do 'Hi' is no longer neither IM nor CM.

However, suppose that only one of the agents take a cooperative stance. If the other does Lo under these circumstances, the best the cooperative agent can do is also to do Lo. But this is not CM. To take a cooperative stance makes no difference in this case. The theory could be adjusted so as to assign taking a cooperative stance some final value in itself, such that it becomes the uniquely best answer for the agent. But the theory is still not CM. Hence, no IM theory, even if it is not exclusively act-oriented by allowing for unilaterally taking a cooperative stance, can be CM in all cases.

Suppose finally, by assuming that it is possible for an agent to successfully influence the other. The pattern (Hi, Hi) resulting from both taking a cooperative stance would be both IM and CM. But again, this only holds if taking a cooperative stance involves no cost. Collective maximization is satisfied only in the pattern (Hi, Hi) resulting from *not* taking a cooperative stance, in which individual maximization is also universally satisfied.

Going for a SCM theory rather than a merely CM one is to accept in cases of disconnection there is reason to make the best of the situation with those who are connected. This may be all right.

The lesson we draw is that a moral solution to a moral Hi-Lo problem is effective only if all agents become convinced and motivated through calls for coordinated action. But an attempt to convince others is likely to involve costs, whereby the coordinated action is no longer guaranteed to be collectively maximizing. Hence, it is not easy to have to both ways even if we reframe the original Hi-Lo case.

References

Bacharach, Michael (2006). Beyond Individual Choice: Teams and Frames in Game Theory. Princeton University Press.

Bykvist, Krister (forthcoming). High stake coordination problems – Do we need to reach beyond individual duties to solve them? In Säde Hormio & Bill Wringe (Eds.), Collective Responsibility: Perspectives on Political Philosophy from Social Ontology. Springer.

Chander, P. (2018) Game Theory and Climate Change. Columbia University Press.

Collins, S. (2019). Collectives' Duties and Collectivization Duties. Australian Journal of Philosophy, 91(2), 231-248.

Cripps, Elizabeth (2013). Climate Change and the Moral Agent: Individual Duties in an Interdependent World. Oxford University Press.

Feldman, Fred (1986). Doing the Best We Can: An Essay in Informal Deontic Logic. D. Reidel Publishing Company.

Gibbard, Allan F. (1965). Rule-utilitarianism: Merely an illusory alternative? Australasian Journal of Philosophy, 43 (2), 211-220.

Goodin, Robert (2012). Excused by the willingness of others? Analysis 71(1), 18-24.

Magli, A. C. & Manfredi, P. (2022). Coordination games vs prisoner's dilemma in sustainability games: A critique of recent contributions and a discussion of policy implications. Ecological Economics 192, 1-7.

Lewis, David Kellogg (1969). Convention: A Philosophical Study. Cambridge, MA, USA: Wiley-Blackwell.

Luce, Robert Duncan & Raiffa, Howard (1957). Games and Decisions: Introduction and Critical Survey. New York: Wiley.

Pellikaan H. & R. J. van der Veer (2002). Environmental Dilemmas and Policy Design. Cambridge: Cambridge University Press.

Pinkert, Felix (2015). What If I Cannot Make a Difference (and Know It). Ethics 125 (4), 971-998.

Portmore, Douglas W. (2018). Maximalism and Moral Harmony. Philosophy and Phenomenological Research (2), 318-341.

Rabinowicz, Wlodzimierz (1989). Act-utilitarian prisoner's dilemmas. Theoria 55, 1-43.

Regan, Donald (1980). Utilitarianism and co-operation. New York: Oxford University Press.

Sugden, Robert (2015). Team Reasoning and Intentional Cooperation for Mutual Benefit. Journal of Social Ontology 1 (1), 143–166.

Schwenkenbecher, A. (2021). Getting our act together. Routledge.

Zimmerman, Michael J. (1996). The Concept of Moral Obligation. New York: Cambridge University Press.

Tim Campbell [1] & Patrick Kaczmarek[2]

Improving Lives and Avoiding Harm: A Critical Response to Harm-Based Arguments for Climate Anti-Natalism[3]

Creating a new person produces more CO2 emissions than many other lifestyle choices, such as driving a gas-powered car, eating meat, and flying. According to Climate Anti-Natalism, for this reason, in many instances, it is wrong to create a new person, even if that person would have a good life. Arguments for Climate Anti-Natalism point to the harm that CO2 emissions cause, but they do not recognize any moral reason to create people with good lives. We identify a harm-avoidance principle underlying arguments for Climate Anti-Natalism. We then show that any moral theory that accommodates this harm-avoidance principle has implausible implications. Such a theory either permits agents to create people with bad lives rather than with good lives, requires agents to harm people just to avoid imposing less harm on those same people, or permits agents to impose any amount of uncompensated harm. A reasonable response to this problem is to reject the harm-avoidance principle, thus undermining the case for Climate Anti-Natalism.

[1] Institute for Futures Studies, Stockholm and Mimir Center for Long Term Futures Research, Stockholm, timothy.campbell@iffs.se.

[2] Centre for the Study of Existential Risk, Cambridge, pakazmarek@gmail.com.

[3] Funding from Effective Giving and Riksbankens Jubileumsfond (grant number: P22-0662) is gratefully acknowledged.

1. Introduction

Empirical evidence suggests having a child generates more CO2 emissions than driving a gas-powered car, flying, and eating meat (Wynes and Nicholas 2017). At the same time, several ethicists have argued that producing large quantities of CO2 by having children is, morally speaking, no different from producing comparable quantities of CO2 by such other means (Young 2001; MacIver 2015; Conly 2015; Hedberg 2019, Rieder 2018; Burkett 2021). Some have gone as far as to claim that for many people in rich countries, having children is morally wrong, since the CO2 emissions from having children contribute to the harm of climate change and are inessential for a decent life (Burkett 2021). Let us call the claim that it is wrong to have children because it would contribute to the harm of climate change *climate anti-natalism*.

The arguments for climate anti-natalism neglect some important questions. This paper focuses on the most central neglected question: can the fact that a person would exist with a good life, by itself, provide a moral reason to create this person? Once this question is brought into focus, it quickly becomes apparent that climate anti-natalists have no easy answer to it. For grappling with the question requires us to dive into the thorny field of normative population ethics, in which every theory is deeply problematic. The theories most naturally suited to climate anti-natalism are those built around the idea that while we have moral obligations to avoid harming people, we have either no moral reason whatsoever, or at least no requiring moral reason, to create people with good lives.

The debate around climate anti-natalism is therefore related to debates in population ethics about what has become known as *the Asymmetry*. In its *moral reasons* variation, the *Asymmetry* states that, when other things are equal, we have moral reason to avoid creating people with bad lives, but no moral reason to create people with good lives when the alternative is to create no one. The *Asymmetry* also has a deontic formulation, according to which, when other things are equal, we are morally required not to create people with bad lives, but not to create people with good lives when the alternative is to create no one. Underlying the deontic version of the *Asymmetry* is a claim that we call *harm-avoidance*: it can be morally impermissible to harm individuals, but refraining from creating individuals with good lives is morally permissible, other things being equal. More generally, some have claimed that in a certain restricted class of choice situations, which we identify in §2.1, an option is impermissible only if it does harm. Call this *the harm-avoidance account* of the Asymmetry.

Several moral theories are based on the harm-avoidance account. Call any such theory a *harm-avoidance theory*. As many have acknowledged, the simplest harm-avoidance theories, those which morally require agents to minimize total harm, face an

especially troubling problem, *the Problem of Improvable Life Avoidance*, which we present in §3.

The search for a harm-avoidance theory that deals adequately with *the Problem of Improvable Life Avoidance* is underway. In this paper, we argue that this search is ill-starred. Any harm-avoidance theory either faces some no less troubling variant of *the Problem of Improvable Life Avoidance*, permits agents to create people with miserable lives rather than with good lives, or permits inflicting any amount of harm, thus giving inadequate consideration to harm-avoidance.

We begin, in §2, by making 'harm-avoidance' more precise, explaining in greater detail what we take to count as a harm-avoidance theory, and identifying a defining feature of any such theory, a commitment that we call *Harmless Permission* (cf. §2.4). In §3, we present *the Problem of Improvable Life Avoidance*. In §§4–5, we scrutinize four state-of-the-art harm-avoidance theories as case studies, specifically those recently put forward by Michael McDermott, Joe Horton, Teruji Thomas, and Abelard Podgorski. These theories seem to offer an adequate response to *the Problem of Improvable Life Avoidance*, but they encounter other, no less severe, problems. By exposing the problems these theories face, we identify three plausible principles: *Weaker Dominance Addition* (cf. §4), *Weak Improvable Life Acceptance* (cf. §5.1), and *Limit Permissible Harm* (cf. §5.2). In §6, we prove that no harm-avoidance theory can satisfy all three principles. Finally, in §7, we conclude with some reflections on what this means for climate anti-natalism. Our result casts doubt on the harm-avoidance account, and points toward the existence of moral requirements to create people with good lives. In other words, our result undermines the case for climate anti-natalism. While it is still possible to defend climate anti-natalism on grounds other than harm-avoidance, a convincing alternative case is yet to be made.

2. The Harm-Avoidance Account and General Theories

According to the harm-avoidance account, in certain choice situations, an option is impermissible only if it does harm. In this section, we spell out what the relevant choice situations are, and what counts as a harm-avoidance theory.

There are four points of clarification regarding harm-avoidance theories: their domain of application (§2.1); their characterization of 'harm' (§2.2); the various conceptual framings they can adopt (§2.3); and what they take to be the moral significance of creating well-off people (§2.4). Regarding the final point, according to our classification, a harm-avoidance theory can recognize a special exception to the rule that harmless options are permissible, namely an exception for *non-identity cases*, or

cases in which the agent can create a new person (or people) but can determine which new person (or people) will exist.

2.1 Domain restriction

A theory based on the Asymmetry *could* be a fully general theory of the permissibility of options.[4] However, most theories we will discuss are not clearly intended to be fully general. Their proponents either explicitly assume, or are most charitably interpreted as assuming, that the application of the theories they defend is restricted to *normative population ethics*—the part of moral philosophy concerned with the permissibility of options that may affect the (i) number, (ii) identities, and (iii) well-being levels of people, and where permissibility facts supervene on facts about (i)–(iii). Considerations other than (i)–(iii), such as personal virtue, agent-relative prerogatives, special obligations, and whether agents lie, cheat, and steal (without affecting the number, identities, and well-being levels of people) are typically sidelined.[5] Among the theories we discuss in §4–5, there is one minor exception. Joe Horton (2021) defends a theory, which we shall include in our classification of harm-avoidance theories, according to which a necessary condition for an option to be impermissible is that it affects a *non-consenting* individual. But this minor exception aside, harm-avoidance theories are concerned only with how considerations (i)–(iii) affect permissibility.

2.2 Harm

The second point of clarification concerns 'harm'. Harm-avoidance theories recognize only two types of harm as morally significant: comparative harm and existential harm.[6]

Suppose your only options are A and B. If a certain person exists given the choice of either A or B, and is worse off given the choice of A, then A comparatively harms her. If she has a bad life given A but does not exist given B, then A existentially harms

[4] See, e.g., Bader 2022b; Cusbert & Kath 2018; McDermott 1982.

[5] Thomas (2022) explicitly assumes a domain restriction along these lines; ? do not mention the topic of domain restriction, but we will charitably interpret them as including it. ? seems to be assuming a restriction of this kind in his defense of the claim that an option is impermissible only if it harms someone.

[6] There are at least two other types of harm discussed in the literature on harm, but they are not relevant to our discussion. First, an option might impose non-comparative harm on an individual by creating her in an intrinsically bad state, or by creating her in a state which has an intrinsically bad aspect (Harman 2009; Shiffrin 1999). Second, an option might impose harm on an individual by making her worse off than she could have been in some specified possible outcome, where this outcome need not be one of the agent's options . See for a discussion of different types of harm.

her.[7] The harm done by a chosen option is defined as comparative or existential relative to some alternative that the agent could choose. It is therefore possible that an option A comparatively harms someone relative to some alternative B, but existentially harms the same person relative to some third alternative, C.

For instance, suppose I face the decision represented in Table 1. The lifetime well-being levels of people affected by my choice are represented numerically in the tables, where positive and negative numbers represent, respectively, positive and negative well-being, 0 represents neutral well-being, and 'Ω' represents non-existence.[8]

Table 1. Improvable Life

	Pebbles
Option 1	Ω
Option 2	−10
Option 3	10

In *Improvable Life*, Pebbles is existentially harmed by Option 2 relative to Option 1, since, given Option 2, Pebbles has a bad life, and she does not exist given Option 1. But Pebbles is comparatively harmed by Option 2 relative to Option 3, since, given Option 2, she is worse off than she is given Option 3. Following Jacob Ross, we will say that Pebbles has an *improvable life* in the outcome of Option 2. A person has an improvable life given the choice of some option, if the chosen option comparatively harms her.

How much harm should we say Option 2 does in Improvable Life? One possibility is that the harm of an option has a certain magnitude relative to some alternative, but there is no such thing as the magnitude of the option's harm *full stop*. For instance, one might claim that Option 2 does existential harm of magnitude 10 relative to Option 1, and comparative harm of magnitude 20 relative to Option 3, but there are no further facts regarding how much harm Option 2 does. The theories considered in §5 are of this sort; they assume the magnitude of any morally significant harm is determined only relative to some alternative.[9]

Another possibility is that there is such a thing as the magnitude of an option's harm full stop. For instance, we could say that the magnitude of any comparative harm full stop is the difference between the harmed individual's well-being in the

[7] On this distinction, see Bykvist (2006); McMahan (1981, 2013), Parfit (2017), Podgorski (2023), and Thomas (2022).

[8] Since at least Boonin (2014), characters from Hanna-Barbera cartoons have sometimes featured in debates in population ethics. We continue the tradition here.

[9] The qualification 'morally significant' is important. Technically, the theories considered in §5 leave open whether there is harm full stop; but if there is, then it is not morally significant on these theories.

outcome in which she is comparatively harmed and her well-being in the outcome in which, among the outcomes the agent can bring about, she is best off. On the other hand, if the only outcome in which the harmed person does not have negative well-being is an outcome in which she does not exist, then the magnitude of the existential harm imposed on her is simply her negative well-being level. Those theories considered in §§3–4 are of this sort.

On any harm-avoidance theory, the total harm of an option is the sum of all individual comparative and existential harms done by the choice of that option; this is either total harm full stop or total harm relative to some alternative, depending on one's theory.

Finally, we acknowledge that the conception of 'harming' employed in this paper is somewhat non-standard, and may strike some readers as odd. There is an understanding of 'harm', as a verb, according to which whether the choice of some option *harms* someone depends on whether the option involves "doing" as opposed to a merely "allowing". According to this understanding, if I choose an option that merely allows a certain person to drown, then I have allowed a morally significant harm, but I have not *harmed* the person, since I did not directly cause her drowning.[10] Given our terminology, however, I *do* harm this person by allowing her to drown. Specifically, assuming she would have been better off had I done something other than allow her to drown, I comparatively harm her. This non-standard use of the active verb 'harm' serves mainly to *abbreviate* discussions of the cases we are interested in. Readers who find it odd that a mere allowing *harms* someone should feel free to interpret sentences such as 'I harmed the person by letting her drown' in some other way, such as 'By letting the person drown, I brought about an outcome in which she was harmed'. What matters for our purposes is that some options for an agent result in harm that certain alternatives for that same agent avoid. It doesn't matter *substantively* whether we describe these options as *harming*. Indeed, the moral relevance of the doing-allowing distinction is yet another deontic consideration that is typically bracketed in discussions of normative population ethics.

2.3 Conceptual framing

Harm-avoidance theories are often couched in terms of 'complaints', or 'objections' on behalf of a person who suffers harm as a result of the agent's choice.[11] One reason for this is related to our last point in the previous sub-section, i.e., the apparent oddness of describing mere allowings as harming. Rather than say that the choice of

[10] Or so many assume, such as F. M. Kamm (1996); cf. Kagan 1991; Otsuka 1997.

[11] While harms may not be the only source of complaints, harms are a necessary source of complaints on the theories we discuss.

some option harms a person, when the option is a mere allowing, some prefer to say that the choice of that option gives the person who thereby suffers harm a complaint against the agent, or an objection to how the agent has behaved, where the basis of this complaint or objection is the harm that this person thereby suffers.

Another reason for adopting the language of 'complaints' or 'objections' is that, as remarks, this language pairs well with certain general moral theories, such as T. M. Scanlon's contractualism, as a matter of what people owe to each other.[12]

Couching a harm-avoidance theory in terms of complaints or objections also allows for an explanation of the Asymmetry that maps neatly onto the harm-avoidance account. Someone who is made to exist with a miserable life when the agent could have refrained from creating them has been existentially harmed, and therefore has a complaint. However, if the agent chooses not to create a person with a good life, then assuming no one else is harmed by the agent's choice, there is no one who can reasonably complain.

2.4 The moral significance of existential benefits

The final point of clarification concerns what a harm-avoidance theory takes to be the moral significance of creating people with good lives, i.e., the significance of conferring *existential benefits*. In a choice between options A and B, A existentially benefits someone if A causes her to exist with a good life but she would not exist given B. Like existential harm, the existential benefit of an option is defined relative to an alternative. The magnitude of an existential benefit to a person is simply her positive well-being level.

All harm-avoidance theories agree that agents have no moral obligation to create existential benefits rather than not create them, all else being equal. However, some harm-avoidance theories entail that creating existential benefits can justify harm, making certain otherwise impermissible harmful options permissible.[13] For instance, it seems permissible to prolong the human race even though at least some future people will have bad lives, and hence, will suffer existential harm. Similarly, it seems permissible for parents to sacrifice some of their well-being for the sake of creating a happy child. One way to capture these intuitions is to claim that conferring large enough existential benefits can justify imposing harm.[14]

---

[12] The focus on complaints also helps bring out one implausibility of *the Problem of Improvable Life Avoidance*: what Horton (2021) refers to as 'backfiring complaints' (cf. §3, below).

[13] Cf. McMahan, 2013; Thomas 2022; Podgorski 2023.

[14] See Thomas (2022, §§4.2-4.3 , as well as [removed for blind review], for discussion of these cases. See also Horton (2021, §1.2 for a discussion of how the claim that creating existential benefits can justify harm avoids what he calls 'the Problem of Tyrannical Complaints'.

Finally, although all harm-avoidance theories say there is no requirement to create existential benefits rather than not create them, as we mention above, a harm-avoidance theory can recognize a requirement to create some people with good lives rather than other people with good lives. See Table 2.

Table 2. Non-identity Case

	Elroy	Judy
Option 1	1	Ω
Option 2	Ω	10

We classify non-identity cases as those in which an agent has at least two options, A and B, where A would create someone who would not exist given B, and B would create someone who would not exist given A. These cases involve a choice between creating different *contingent* people, i.e., people whose existence depends on the agent's choice. In contrast, *addition cases* are those in which an agent can choose whether to create some contingent person (or people), but cannot choose between creating *different* contingent people.

In *Non-identity Case*, you can either existentially benefit one contingent person (Elroy) or existentially benefit a different contingent person (Judy) even more. It seems that creating Elroy (Option 1) does not harm Judy, and creating Judy (Option 2) does not harm Elroy.[15] Since neither option does harm, many who defend what we classify as a harm-avoidance theory would accept the claim that either option is permissible. They would therefore reject the following intuitively plausible principle:

> **Normative Egalitarian Dominance (NED):** For any options A and B, if the population that exists given the choice of A has perfect equality of welfare, is the same size as the population that exists given B, and every person who exists given A has higher welfare than every person who exists given B, then B is impermissible, other things being equal.[16]

However, our classification of harm-avoidance theories is broad enough to include theories that accommodate NED. In a choice between A and B, where A creates some contingent person S who would not exist given B, and B creates some contingent person S* who would not exist given A, let us say that A creates a *non-identity shortfall*

[15] But see Meacham (2012) for an opposing view.

[16] The title of this principle is originally due to Arrhenius (2022, p. 191). Our statement of the principle is similar to Arrhenius's, except that in ours, 'A' and 'B' denote possible options for an agent rather than populations, and we use the term 'impermissible' rather than 'wrong'.

if A existentially benefits S less than B existentially benefits S*. On our broad classification, a harm-avoidance theory can recognize both non-identity shortfall and harm imposition as sources of impermissibility. For instance, Thomas (2022) presents such a theory, which is compatible with NED (cf. §5.1). In a similar vein, Johann Frick and Michael Otsuka each propose a set of principles that reconciles the Asymmetry and NED.[17]

A harm-avoidance theory, then, is any theory that includes the following commitment:

> **Harmless Permission:** If option A does no comparative or existential harm, and does not create any non-identity shortfall, then A is permissible.

Where the only considerations assumed to be relevant to permissibility are the number, identities, and well-being levels of people.

3. The Problem of Improvable Life Avoidance

Harmless Permission is compatible with a wide range of different general theories. The simplest harm-avoidance theory, which often serves as a starting point in discussions of how best to develop a harm-avoidance theory, is

> **Harm Minimization:** An option A is permissible iff there is no alternative B which does less total harm than A.

The magnitude of a comparative harm is here assumed to be the difference between the harmed person's well-being in the outcome in which she is harmed and her well-being in the outcome in which, among the outcomes the agent could have brought about, she is best off. And the magnitude of an existential harm is assumed to be the harmed person's negative well-being level, where the only alternatives to existentially harming the person involve not creating her at all.

Harm Minimization implies that it is impermissible to create a miserable person rather than not create them, other things being equal. Creating the person would impose some existential harm, but refraining from creating the person would impose no harm, so not creating the person would do less harm than creating them. Harm Minimization also implies that it is permissible not to create a happy person rather

[17] Frick (2020) defends the moral reasons formulation of the Asymmetry, not the deontic formulation. However, the moral reasons formulation supports the deontic formulation, and Frick presumably accepts the latter in addition to the former.

than create them, other things being equal. In this case, not creating the happy person would do no harm, so there can be no alternative that does even less harm.

However, as several philosophers have pointed out, Harm Minimization faces serious problems, one of which is *the Problem of Improvable Life Avoidance*.[18] Table 3 illustrates a case, due to Jacob, which is commonly used to introduce the problem.[19]

Table 3. Ross's Case

	Roxy	Chip
Option 1	1	Ω
Option 2	10	10
Option 3	−2000	1000

In *Ross's Case*, Option 1 imposes a comparative harm of 9 on Roxy, Option 2 imposes a comparative harm of 990 on Chip, and Option 3 imposes a comparative harm of 2010 on Roxy. Option 1 minimizes total harm. Harm Minimization therefore implies that Options 2 and 3 are impermissible, and that Option 1 is morally required, as it is the only permissible option.

But this assignment of deontic statuses to options 1–3 seems implausible. Although Option 3 is clearly impermissible, the claim that Option 2 is also impermissible, and that Option 1 is therefore morally required, is problematic for at least two reasons.

First, it implies that the agent is morally required to avoid creating a certain person with a good life just because this life would be improvable. This is where *the Problem of Improvable Life Avoidance* gets its name. If Option 2 is impermissible, this can only be because it gives Chip an improvable life. Yet, Option 2 also gives Chip a *good* life. If Option 2 is impermissible only because it gives Chip a good but improvable life, it may seem odd that one is morally required not to create Chip. Presumably, Chip should prefer existence with a good life to non-existence. For instance, if we choose Option 2, and Chip objects that we have harmed him, we can respond "the only alternative for us that would not inflict even greater harm on someone else would leave you out of existence altogether. Is that really what you want?" One imagines that his answer would be "No".[20]

Notice also that according to Harm Minimization, in a binary choice between Options 1 and 2, Option 2 minimizes harm. So in this binary choice, Option 2 is morally required, and hence permissible, whereas Option 1 is impermissible, and hence, *not*

[18] See Thomas (2022, §2) for a full discussion of the difficulties with Harm Minimization.

[19] Our presentation of the case was sourced, with minor cosmetic changes, from Podgorski (2023, p. 353).

[20] This is an instance of what Horton dubs a 'backfiring objection' (cf. footnote 11). See also McDermott (2019).

morally required. But on Harm Minimization, Option 2 *becomes* impermissible, and Option 1 *becomes* morally required, when the horrible Option 3 is added to the option set. Let us say that a moral consideration against some option is a possible source of that option being impermissible. Then Harm Minimization violates

> **Improvable Life Acceptance (ILA):** If (i) person S has a good life given A, (ii) does not exist given B, and (iii) the only moral consideration against A in a choice from some option set O that includes A and B is that A comparatively harms S, then if B is not morally required in a binary choice between A and B, then B is not morally required in a choice from O.

Basically, ILA says that just to avoid giving someone a good but improvable life, one is not morally required to leave that person out of existence.[21]

While ILA strikes us as fairly plausible, some would reject it on the grounds that in certain cases, refraining from creating a person with a good but improvable life is the only way to avoid unjust harm.[22] Perhaps one reason why Harm Minimization's violation of ILA seems implausible when considering *Ross's Case* is that in this case the only option that gives Chip a better life than Option 2 is Option 3, which is unspeakably horrible for Roxy, and is, as Otsuka (2017) would say, "manifestly unreasonable". One might think that it is this detail, and not a principle such as ILA, that explains why the comparative harm that Option 2 imposes on Chip is insufficient to make Option 2 impermissible in a choice between Options 1–3.

Whatever one thinks about ILA, *Ross's Case* illustrates a second problem with Harm Minimization, namely that it morally requires dominated options. One option dominates another *iff* it is better for someone and worse for no one. In *Ross's Case*, Option 2 dominates Option 1, since it is better for Roxy and worse for no one. In fact, Option 2 *weakly addition-dominates* Option 1. Option A *addition-dominates* option B *iff* (i) everyone who exists given B would be better off if A were chosen, (ii) A creates

---

[21] The following statement by Podgorski comes close to capturing the core idea of ILA: "It should not be possible to start with a set of choices which permit us to create someone with a happy life, add an option under which they are better off, and thereby generate a complaint *on their behalf* which makes it impermissible to create them at all" (2023, p. 354). There are only two differences between Podgorski's statement and our formulation of ILA; first, Podgorski refers to a "complaint" on behalf of the person mentioned, and second, he speaks of the option of creating the person being 'permitted' or 'impermissible', whereas we speak of the option of *not* creating the person as being 'not morally required'. Our formulation is weaker than Podgorski's, since it doesn't assume anything about complaints, and it leaves open the (admittedly implausible) possibility of the agent facing a moral dilemma when C is added to the option set alongside A and B.

[22] See, e.g., Boonin (1996) and Frick (2022), as well as Temkin (2012, ch. 13). Boonin and Frick discuss Parfit's *Mere Addition Paradox* as a case where adding people with good but improvable lives results in a morally worse outcome relative to the option set. Ingmar Persson (2017, ch. 8) argues that it can be worse to add people with good but improvable lives if this increases inequality; and Temkin (2012, ch. 12) suggests that this may be the case.

some people who would not exist given B, and (iii) everyone who exists given A has a good life. On the other hand, A *weakly* addition-dominates B *iff* A addition-dominates B and everyone who exists given A has equal well-being. Since Harm Minimization implies that Option 1 is morally required in *Ross's Case*, it violates

> **Weak Dominance Addition Exemption** (**WDAE**)**:** If option A weakly addition-dominates option B, then B is not morally required.[23]

Most harm-avoidance theorists seem to agree we should accept WDAE.[24]

We have two problems here for those who wish to develop a general harm-avoidance theory. First, the simplest harm-avoidance theory violates ILA (the Problem of Improvable Life Avoidance); second, it violates WDAE. For harm-avoidance theorists, there are different ways of responding to these problems. They could find a compelling justification for rejecting ILA and WDAE, and perhaps supply an alternative explanation of where Harm Minimization goes wrong in *Ross's Case*, such as that comparative harm doesn't count against an option when the alternative that is better for the harmed person is "manifestly unreasonable". They could formulate a harm-avoidance theory that satisfies both ILA and WDAE without violating some equally compelling principle. Or they could adopt a mixed approach that involves formulating a theory that accommodates only one of the two principles, while supplying a justification for rejecting the other. McDermott (2019), Horton (2021), and Podgorski (2023) have taken the second approach, while Thomas (2022) adopts a combination of the second and third approaches, offering one theory that satisfies WDAE and ILA, and a second theory that satisfies WDAE but not ILA.

However, as we will now argue, these harm-avoidance theories have other problematic implications, some even more implausible than violating WDAE, and others even more implausible than violating ILA.

---

[23] Our formulation of this principle is inspired by Elliot Thornley, who appeals to a weaker principle in arguing against a theory defended by Joe Horton, which we consider in §3. Thornley (2023, p. 522) calls his principle 'Weak Normative Dominance Addition', which is like WDAE, except it says that if everyone has non-negative well-being in the weakly addition-dominated option, then if that option is permissible, the weakly addition-dominating option is also permissible.

[24] Some seem to believe that dominated options can be required. For example, Frick (2022, 238ff) suggests that in one "supercharged" version of *the Mere Addition Paradox*, where one of the options is dominated, the dominated option might be morally required. Frick explicitly argues only for the claim that the dominated option is better than the dominating option relative to a certain set of options; but his discussion suggests that the dominated option is also the best option in this set, and that given the absence of any non-axiological considerations, it is required.

4. Avoid Reasonable Objections

McDermott (2019) and Horton (2021) each defend a harm-avoidance theory that accommodates both ILA and WDAE. Although McDermott's and Horton's theories differ in important ways, both say that an option is permissible *iff* no one can reasonably object to it, where a necessary condition for someone reasonably objecting to an option is that they would be harmed by it.

On McDermott's theory, which he calls *'Objection Minimization'*, an individual S has a reasonable objection to an option A *iff* S exists given A, and there is some alternative B such that (i) B is better for S than A (and hence, A harms S), and (ii) B does less total harm than A. Objection Minimization implies that in *Ross's Case* Chip does not have a reasonable objection to Option 2, since the only option that is better for Chip, Option 3, does more total harm than Option 2. It also implies that Roxy does not have a reasonable objection to Option 2, since there is no alternative that is better for her. So according to Objection Minimization, Option 2 is permissible.

One drawback of Objection Minimization is that it implies that Option 1 is permissible.[25] Since Option 2 does more total harm than Option 1, Objection Minimization implies that Roxy cannot reasonably object to Option 1. Since Roxy is the only potential objector to Option 1, no one can reasonably object to Option 1. So Objection Minimization violates

> **Weak Dominance Addition** (**WDA**)**:** If A weakly addition-dominates option B, then B is impermissible.

However, permitting weakly addition-dominated options is not as implausible as requiring them. So Objection Minimization seems at least to improve upon Harm Minimization.[26]

Horton's harm-avoidance theory implies that in *Ross's Case* Option 2 is morally required and Options 1 and 3 are impermissible, which seems correct. Horton's formulation of his criteria for reasonable objectionableness are somewhat complicated. For ease of exposition, our statement of the criteria differs slightly from Horton's, but this doesn't affect our arguments.[27]

According to Horton's *Avoid Reasonable Objections*:

[25]Thomas (2022, fn 8) makes this point.

[26] Cf. Thomas 2022, fn 23.

[27] His original statement can be found at (Horton 2021, p. 499).

A person can reasonably object to an option A *iff* she exists, has not consented to A, and there is some alternative option B satisfying 1—4.[28]

1. B is better for her than A.
2. B gives a greater sum of well-being than A to the set of people who exist given A.
3. The sum of well-being that B gives to the set of people who exist given B is greater than the sum of well-being that A gives to the set of people who exist given A.[29]
4. No one can reasonably object to B.

An option is permissible *iff* no one can reasonably object to it.

Two clarifications are needed. First, Horton thinks that B can be better or worse for someone than A, even if she does not exist given B. If she has a bad life given A, but does not exist given B, then on Horton's view, assuming she exists (i.e., A has been chosen), B is better for her than A. Second, for the purpose of determining whether condition 2 is satisfied, the sum of well-being that *B* gives to the set of people who exist given A is the sum of the individual well-being values *of B* for those who exist given A, where some people who exist given A may not exist given B. If someone who exists given A does not exist given B, then, Horton assumes, B gives *zero* well-being to this person.

According to Avoid Reasonable Objections, in *Ross's Case*, Option 2 is morally required because it is the only option no one can reasonably object to. To see this, notice that neither Roxy nor Chip can reasonably object to Option 2. Chip cannot reasonably object to Option 2 because his objection cannot satisfy conditions 2 and 3. The only option that is better for Chip than Option 2 is Option 3, which produces less well-being than Option 2 for the set {Roxy, Chip}. Roxy cannot reasonably object to Option 2 because her objection cannot satisfy condition 1, i.e., there is no alternative to Option 2 that is better for Roxy.

[28] Although it may be unclear from the four conditions stated here, Avoid Reasonable Objections is indeed a harm-avoidance theory. According to this theory, the permissibility of an option requires that no one could reasonably object to it, and a necessary condition for reasonably objecting to an option is that there is some alternative that is better for the objector. Where A is the option to which the objector objects, if the relevant alternative B is better for the objector than A because the objector has a bad life given A and does not exist given B, then the objector is existentially harmed by A; on the other hand, if B is better than A for the objector because she would exist at a higher level of well-being given B, then the objector is comparatively harmed by A. So on Avoid Reasonable Objections, an individual has a reasonable objection to A only if A harms her.

[29] We've included Thornley's (2023, p. 519) amendment to Horton's condition 3 in our statement.

On the other hand, Roxy can reasonably object to Option 3. Option 2 is better for Roxy than Option 3, so condition 1 is satisfied. Option 2 also produces a greater sum of well-being than Option 3 for the set {Roxy, Chip}. This is sufficient for Roxy's objection to satisfy conditions 2 and 3 because Roxy and Chip are the only people who exist given either Option 2 or Option 3. Finally, because no one can reasonably object to Option 2, Roxy's objection to Option 3 satisfies condition 4.

Roxy can also reasonably object to Option 1, since Option 2 is better for her, gives a greater sum of well-being than Option 1 to the set {Roxy}, the sum of well-being that Option 2 gives to the set {Roxy, Chip} is greater than the sum of well-being that Option 1 gives to the set {Roxy}, and no one can reasonably object to Option 2.

Since someone can reasonably object to Options 1 and 3, and no one can reasonably object to Option 2, Option 2 is the only permissible option, according to Avoid Reasonable Objections.

Although Avoid Reasonable Objections provides a plausible treatment of *Ross's Case*, and does not require weakly addition-dominated options, Horton acknowledges that it sometimes *permits* weakly addition-dominated options, thereby violating WDA. He also acknowledges that Avoid Reasonable Objections sometimes implies that when one option weakly-addition dominates another, the latter is permissible and the former impermissible, an objection emphasized by Thornley (2023).

Horton illustrates this with the following case, which he thinks demonstrates a *strength* of Avoid Reasonable Objections:[30]

Table 4. Horton's Case 6

	Barney	Betty
Option 1	1	Ω
Option 2	2	2
Option 3	Ω	100

In *Horton's Case 6*, Option 2 weakly addition-dominates Option 1. But Avoid Reasonable Objections implies that Option 2 is impermissible and that Option 1 is permissible. Option 2 is impermissible because Betty has a reasonable objection to it. She exists given Option 2, she does not (Horton assumes) consent to this act, and there is an alternative to Option 2, namely Option 3, which is better for Betty, produces more well-being for the set {Betty, Barney}, and produces more well-being for {Betty} than Option 2 produces for {Betty, Barney}. Moreover, no one has a reasonable objection to Option 3 according to Avoid Reasonable Objections, since Betty is the only person who exists given Option 3, and there is no alternative to Option 3 that is better for

[30] Our presentation of his case was sourced, with minor cosmetic changes, from Horton (2021, p. 496).

her. So Betty's objection to Option 2 satisfies all four of Horton's conditions for reasonableness. Because Betty has a reasonable objection to Option 2, Barney *does not* have a reasonable objection to Option 1. The only option that would be better than Option 1 for Barney, namely Option 2, is one that Betty can reasonably object to; so Barney's objection to Option 1 does not satisfy condition 4. Since no one has a reasonable objection to Option 1, according to Avoid Reasonable Objections, Option 1 is permissible in *Horton's Case 6*.

The fact that Avoid Reasonable Objections implies Option 1 is permissible but Option 2 impermissible seems like a problem, though again, this problem is not as grave as that of requiring weakly addition-dominated options.[31] Intuitively, Option 3 is *morally required* in *Horton's Case 6*. Not only does Option 3 produce the most well-being of any option, but more importantly, it is the only option that avoids harm. However, Horton thinks we should reject the claim that Option 3 is morally required because, on Avoid Reasonable Objections, this claim violates ILA. Recall that according to ILA, if, in a choice between creating someone with a good life (A) and leaving them out of existence (B), we are not required to choose B, then adding another option (C) that is better for this person than A cannot generate a moral requirement to choose B. To see why requiring Option 3 would violate this principle on Avoid Reasonable Objections, consider the following. According to Avoid Reasonable Objections, in *Horton's Case 6*, Option 3 would not be morally required in a binary choice between only Options 1 and 3, since, in such a binary choice, neither option would harm anyone. If Option 3 is not required in a binary choice between Options 1 and 3, but Option 3 *becomes* required when Option 2 is added to the option set, then we have a straightforward violation of ILA. Barney exists with a good life given Option 1, he does not exist given Option 3, and the only moral consideration against Option 1 is that it comparatively harms Barney, since Option 2 is better for him than Option 1.

Two points are worth emphasizing here. First, intuitively, Option 3 *is* morally required in a binary choice between Options 1 and 3. In this binary choice, Option 1 creates one person (Barney) rather than a different person (Betty) with a much better

[31] Note that Avoid Reasonable Objections violates Thornley's Weak Normative Dominance Addition principle. It is worth emphasizing that in *Horton's Case 6*, the possible well-being levels for those who might exist given the different options seem to be chosen to minimize the intuitive implausibility of Avoid Reasonable Objections's violation of WDA. These levels are chosen so that the difference in well-being for Betty between Options 2 and 3 is relatively large, while the difference for Barney between Options 1 and 2 is relatively small. But as Thornley (2023, p. 522) points out, Avoid Reasonable Objections would assign the same deontic statuses to the three Options in any case with a similar structure but where the well-being difference for Barney between Options 1 and 2 is much greater, and just barely smaller than the well-being difference for Betty between Options 2 and 3. (For instance, imagine, following Thornley's example, that both Barney and Betty would have well-being 49 in the outcome of Option 2.)

life. Since Avoid Reasonable Objections denies this, it violates NED.[32] More generally, it fails to account for the apparent moral significance of non-identity shortfall. Horton would not be moved by this objection, since he bites the bullet in response to *the Non-identity Problem*. However, a harm-avoidance theorist could try to modify Avoid Reasonable Objections to account for the moral significance of non-identity shortfall. For instance, one could keep Horton's criteria for an objection being reasonable, but modify his criterion of permissibility to allow for the possibility of an option being impermissible when it causes non-identity shortfall. Of course, one would then need to figure out how to fit these different criteria together into a coherent moral theory. The theoretical benefit would be that one could say that Option 3 is morally required in *Horton's Case 6* without rejecting ILA.

Second, even if a harm-avoidance theorist insists on rejecting NED, she could reasonably view *Horton's Case 6* as a counterexample to ILA. Even if one claims that Options 1 and 3 are both permissible in a binary choice, one could justify the claim that Option 1 becomes impermissible when Option 2 is added to the option set on the grounds that Option 1 then harms someone. Specifically, one could claim that if an agent must create someone with a good life, and the agent's choice is between creating a person with a good but improvable life and creating a different person with a life that is at least as good but *unimprovable*, then the agent ought to create the latter. Violating ILA in this type of case may not seem too high a cost for securing the plausible judgment that Option 3 is morally required in *Horton's Case 6*.

Unfortunately, regardless of how one addresses the foregoing points, Avoid Reasonable Objections faces a further serious objection. This objection applies to McDermott's Objection Minimization as well. Both theories violate

> **Weaker Dominance Addition** (**Weaker DA**)**:** If A weakly addition-dominates B, and everyone who exists given B has a bad life, then B is impermissible.

Permitting weakly addition-dominated options that give everyone a bad life seems even more implausible than requiring weakly addition-dominated options which, like Option 1 in *Ross's Case*, at least give everyone a good life.

To see that Avoid Reasonable Objections and Objection Minimization violate Weaker DA, consider the following case:

[32] McDermott's Objection Minimization also violates NED, since his theory implies that someone can reasonably object to an option only if it harms her, whereas NED implies that an option is impermissible if it creates a person with less well-being than some other person whom one could instead have created, even if this option does not harm anyone.

Table 5. Weaker Dominance Addition Violation

	Wilma	Fred
Option 1	−100	Ω
Option 2	100	100
Option 3	−100	1000

In this case, everyone who exists given Option 1 has a bad life, everyone who exists given Option 2 has a good life, and Option 2 weakly addition-dominates Option 1. Yet both Avoid Reasonable Objections and Objection Minimization imply that Option 1 is *permissible*.

According to Objection Minimization, Option 2 does more total harm than Option 1; Option 2 imposes a comparative harm of 900 on Fred, while Option 1 imposes a comparative harm of only 200 on Wilma. Moreover, there is no alternative to Option 1 that is both better for Wilma and does less total harm than Option 1. Option 3 imposes a comparative harm of 200 on Wilma, and hence does the same amount of harm as Option 1; moreover Option 3 is not better than Option 1 for Wilma. So according to Objection Minimization, Wilma's objection to Option 1 is not reasonable. Hence, according to Objection Minimization, Option 1 is permissible.

To see how the same problem arises for Horton's theory, notice that according to Avoid Reasonable Objections, no one has a reasonable objection to Option 3. If anyone had a reasonable objection to Option 3, it would be Wilma. But the only alternative that is better than Option 3 for Wilma is Option 2, which produces less well-being than Option 3 for {Wilma, Fred}. So Wilma's objection to Option 3 cannot satisfy conditions 2 and 3. Given that no one has a reasonable objection to Option 3, Fred has a reasonable objection to Option 2; he exists given Option 2, does not (we assume) consent to Option 2, and there is an alternative, Option 3, which is better than Option 2 for Fred, produces more well-being than Option 2 for the set {Wilma, Fred} and produces more well-being for {Wilma, Fred} than Option 2 produces for {Wilma, Fred}. Finally, since Fred has a reasonable objection to Option 2, it follows that Wilma's objection to Option 1 cannot satisfy condition 4. She has no reasonable objection to Option 1, making Option 1 permissible according to Avoid Reasonable Objections.

This problem cannot be avoided by modifying Avoid Reasonable Objections and Objection Minimization to accommodate the apparent moral significance of non-identity shortfall. This is because *Weaker Dominance Addition Violation* is not a non-identity case, but an addition case. There is no pair of options where one involves creating a certain contingent person and the other involves creating a different contingent person.

Horton and McDermott might try to avoid the implausible implication that Option 1 is permissible by adopting a prioritarian weighting that assigns greater moral weight to harm-based objections the worse off the objector is.[33] This would allow one to say that Wilma's objection to Option 1 carries more weight than Fred's objection to Option 2. However, this move won't help, since we can imagine the well-being difference for Fred between Options 1 and 2 to be as great as we want. Unless some harm-based objections are infinitely more morally weighty than others, where both objections are based on finite amounts of harm, there will be some magnitude of harm that we can imagine Option 2 imposing on Fred such that his objection to Option 2 will outweigh Wilma's objection to Option 1. We can also imagine a case with the same structure as *Weak Dominance Addition Violation* in which *any number of people* in Fred's position would be harmed by Option 2 to the same extent as Fred. It is hard to see how Wilma's harm-based objection to Option 1 can outweigh any number of Fred-like objections to Option 2.

A different response on behalf of Horton's and McDermott's theories would be to modify these theories to allow for the possibility that someone can have a reasonable objection to an option even when the only alternative that is better for them does more total harm, or is such that someone else can reasonably object to it. This would allow that Wilma's objection to Option 1 is reasonable even though Fred has a reasonable objection to Wilma's preferred alternative, Option 2. However, modifying either theory in this way would give up on its treatment of *Ross's Case*. Specifically, it would open the door to the possibility that in *Ross's Case*, each option is one that someone can reasonably object to. This would make *Ross's Case* a moral dilemma, which seems absurd.

Finally, we note that Harm Minimization also implies that Option 1 is permissible in *Weaker Dominance Addition Violation*, since no alternative to Option 1 does less harm than Option 1. Thus, all three theories we've considered so far violate Weaker DA, which is, we think, worse than violating WDAE.

5. Tournament Theories

Unlike Harm Minimization, Objection Minimization, and Avoid Reasonable Objections, the two harm-avoidance theories considered in this section satisfy WDAE, WDA, and Weaker DA. They forbit weakly addition-dominated options.

Both Thomas's and Podgorski's theories adopt what Podgorski calls a *tournament approach*, and we shall refer to them as 'tournament theories'. A tournament theory is

[33] Horton (2021, n. 6, 17) considers this kind of view.

structured in two parts. First, it includes a set of conditions for when one option defeats another option, or when one ought to choose the former over the latter, in a binary choice, or pairwise comparison, of only those two options. Second, it includes a further condition that determines when the choice of some option from any finite set of options is (im)permissible, where this determination is based on how each option fares in pairwise comparisons with the other options.

In contrast to the theories considered in §§3–4, on the tournament approach, what matters is how much harm an option does in a pairwise comparison with each alternative (cf. §2.2). For instance, if our options are A, B, and C, we must ask how much harm A does in a pairwise comparison with B (ignoring the presence of C), then how much harm A does in a pairwise comparison with C (ignoring the presence of B), and then apply the same procedure to Options B and C. Depending on how each option fares in a pairwise comparison with the others, our general criterion of permissibility will then give us the deontic status of each option.

To illustrate, recall *Ross's Case*, in which, according to Harm Minimization, Option 2 does more harm than Option 1, despite the fact that Option 2 weakly addition-dominates Option 1. In contrast with Harm Minimization, a tournament approach says that the harm done by Option 2 is relevant only in a pairwise comparison of Options 2 and 3. Since Option 3 is clearly horrible, any sensible tournament theory will imply that Option 2 defeats Option 3, or that Option 2 ought to be chosen over Option 3, in a binary choice between Options 2 and 3. But notice also that in a binary choice between Options 1 and 2, Option 1 does more harm than Option 2, since, in that binary choice, Option 1 harms Roxy but Option 2 harms no one. Option 2 therefore "wins" in a pairwise comparison with either Option 1 or Option 3. Thus, any sensible tournament theory will imply that Option 2 is at least permissible. In fact, on Thomas's and Podgorski's theories, Option 2 is morally required in *Ross's Case*.

However, as we will now argue, each of these theories encounters problems which stem, at least in part, from the tournament approach.

5.1 The Maximization Theory

The first tournament theory we consider is due to Thomas (2022).

Thomas actually presents two theories that differ regarding their treatment of *the Non-identity Problem*. The first, which Thomas calls 'a narrow theory' rejects NED, biting the bullet in response to *the Non-identity Problem*. The second, which he calls 'a

wide theory', accommodates NED. He considers *the Non-identity Problem* so vexed that he leaves it an open question which of these theories is more plausible.[34]

Each theory includes its own criteria for when one *ought to choose* some option A over another B in a binary choice between A and B. Because the objection that we will raise applies to both the narrow and wide theories, we will here consider only the narrow theory, which is the simpler of the two theories.

The conditions governing pairwise comparisons on the narrow theory are presented as follows. In a binary choice between any options A and B, let Harm(A) be the total harm (both comparative and existential) that arises from choosing A over B. Let ExBen(A) be the total existential benefit in A, and ExBen(B) the total existential benefit in B. Then one ought to choose A over B *iff*:

1. Harm(B) > Harm(A)
2. Harm(B) + ExBen(A) > Harm(A) + ExBen(B)

The motivation for conditions 1 and 2 is as follows. First, we have moral reasons to avoid comparative and existential harm. These reasons have requiring strength proportionate to the magnitude of the harm that would be inflicted. We also have moral reasons to create existential benefits. However, these reasons have *no* requiring strength; they have only justifying strength.[35] They can defuse competing requiring reasons to avoid harm, but they cannot by themselves generate moral requirements. The justifying strength of one's reason to existentially benefit someone is proportionate to the magnitude of the existential benefit.[36] The narrow theory's condition 1 reflects the idea that there is requiring moral reason to avoid harm, and that in a binary choice between two options it is never the case that the agent ought to choose the option that she has more requiring reason not to choose, i.e., the option that does greater harm. Condition 2 reflects the idea that the purely justifying moral reason to existentially benefit people can neutralize the requiring strength of the moral reason to avoid harm, but also that this purely justifying moral reason cannot by itself make it the case that an agent ought to choose one option over another in a binary choice. Notice, for example, that on the narrow theory, it is not the case that one ought to create a person with a very good life rather than some other person with a life that is good, but not very good. In this binary choice, neither option does less harm than the

[34] But see [removed for blind review] for arguments in favour of the narrow theory over the wide theory.

[35] Rebelling against the old fashion that reasons exclusively issue *pro tanto* requirements, philosophers are increasingly adopting the position that reasons can vary on at least two dimensions with respect to their normative strength (Gert 2004; Kaczmarek & Lloyd forthcoming; Kamm 1985; Lazar 2013; Munoz 2021; Pummer 2023). See esp. Little and Macnamara (2021) for an overview of this literature.

[36] Notice Thomas (2022, p. 490) crafted condition 2 to express the plausible idea that non-requiring reasons justify harm only when the *net* existential benefits favour that outcome.

other (since neither does any harm) and so neither satisfies the narrow theory's condition 1.

Thomas's criterion of permissibility for the narrow (as well as the wide) theory is:

> **Maximization:** In a choice between finitely many options, all and only the maximal options are permissible.

That an option is 'maximal' means that *it is not the case* that one ought to choose some other option over it in a binary choice. Hence, whether A is permissible in a choice between finitely many options, depends on whether there is some B in the option set such that one ought to choose B over A in a binary choice. If so, then A is impermissible. Otherwise, A is permissible.

The narrow theory is the conjunction of Maximization and conditions 1 and 2. To see that the narrow theory satisfies WDA, consider any choice context in which A and B are options. If A weakly addition-dominates B, then the only difference between A and B, when compared pairwise, is that everyone who exists given B also exists given A with higher (positive) welfare, and A creates some additional people, also with positive welfare, who do not exist given B, such that everyone who exists given A is equally well-off. It follows that in a binary choice between A and B, the choice of B would impose some (comparative) harm, but would not create any existential benefits, while the choice of A would impose no harm and would create some existential benefits. Hence, when A weakly addition-dominates B, $Harm(B) > Harm(A)$, and $Harm(B) + ExBen(A) > Harm(A) + ExBen(B)$. So according to the narrow theory's conditions 1 and 2, one ought to choose A over B. Since one ought to choose A over B, it follows from Maximization that in any choice context that includes A and B, the choice of B is impermissible. WDA is satisfied, and since WDA entails WDAE and Weaker DA, the latter are also satisfied.

However, as we illustrate below, the narrow theory leads to a troubling form of improvable life avoidance that we call *strong improvable life avoidance*. Although we have not here considered the wide theory, the contexts in which the narrow theory leads to strong improvable life avoidance are those in which the narrow and wide theories agree on which options are (im)permissible. So strong improvable life avoidance is a problem for both theories. Using the label *'the Maximization Theory'* for the disjunction of the wide and narrow theories, our objection to the Maximization Theory is that it entails strong improvable life avoidance.[37]

To illustrate, consider the following case:

[37] 'The Maximization Theory' is our label, not Thomas's.

Table 6. Strong Improvable Life Avoidance

	George	Jane
Option 1	100	Ω
Option 2	101	0
Option 3	−100	202

On the Maximization Theory, the deontic statuses of the options are those given by the narrow theory. They are determined as follows. First, consider a binary choice between Options 1 and 2. In this binary choice, Option 1 does more harm than Option 2. Moreover, since Jane is the only contingent person, and she has welfare 0 (a neutral life) given Option 2, neither option creates any existential benefits. Hence, one ought to choose Option 2 over Option 1.

Next, consider a binary choice between Options 2 and 3. In this binary choice, Option 2 does more harm than Option 3. Specifically, Option 2 imposes harm of 202 on Jane, while Option 3 imposes harm of only 201 on George. Moreover, neither option produces existential benefits, since the same people exist given either option. Hence, one ought to choose Option 3 over Option 2.

Finally, consider a binary choice between Options 1 and 3. Here, according to the Maximization Theory, neither option is such that it ought to be chosen over the other. Although Option 1 does less total harm than Option 3, it does not produce any existential benefits, while Option 3 produces an existential benefit for Jane that is larger than the comparative harm that Option 3 imposes on George. Thus, Option 3 is the only maximal option in this case, i.e., the only option such that no other option ought to be chosen over it in a binary choice. So by Maximization, Option 3 is the only permissible option, and is therefore morally required.

Notice that because neither Option 1 nor Option 3 ought to be chosen over the other in a binary choice, by Maximization, in such a binary choice, Option 1 is permissible and therefore Option 3 is *not* morally required. It is only when we add Option 2, which is (slightly) better for George than Option 1, to the option set, that the Maximization Theory requires Option 3, which is (much) worse for George than either Option 1 or Option 2. The Maximization Theory therefore violates

> **Weak Improvable Life Acceptance** (**WILA**)**:** If (i) A imposes greater harm on person S than B, and (ii) the only moral consideration against B, in a choice from an option set O that includes A and B, is that B harms S, then if A is not morally required in a binary choice between A and B, then A is not morally required in a choice from O.

As in our statement of ILA, here, by 'the only moral consideration against B' we mean the only potential source of B's being impermissible.

The intuitive idea behind WILA can be grasped by first recalling the intuitive idea behind ILA: to avoid giving someone a good but improvable life, one is not morally required to leave her out of existence. In contrast, the idea behind WILA is that to avoid giving someone an improvable life, one is not morally required to instead give her a life that is *even more* improvable.

In the *Strong Improvable Life Avoidance* case, the only morally relevant consideration against Option 1, on the Maximization Theory, is that it comparatively harms George. But how can *that* generate a moral requirement to impose even greater harm on George by choosing Option 3? A requirement to choose Option 3 would be understandable if we had a requiring reason to existentially benefit Jane rather than leave her out of existence. But on the Maximization Theory, as on all harm-avoidance theories, there is no such requiring reason.

To see that WILA is more plausible than ILA, recall that, as we suggested in §2 and §3, someone could reject ILA on the grounds that in certain cases, in order to avoid comparatively harming someone, we can be required not to create them even with a good life. For instance, this might be the only way to avoid unjust harm. But this rationale for rejecting ILA does not support rejecting WILA. It is patently absurd to claim that just to avoid comparatively harming someone, we can be required to comparatively harm that same person *even more*.

The Maximization Theory violates WILA because it rules out options as impermissible on the basis of pairwise comparisons. For instance, Option 1 is deemed impermissible solely on the basis of a pairwise comparison with Option 2, and Option 2 is deemed impermissible solely on the basis of a pairwise comparison with Option 3. The Maximization Theory therefore cannot account for any potentially morally significant relations between Options 1—3 when all three options are considered together, for instance, the fact that the person who would be harmed by Option 1 in relation to Option 2 (George) is the same person who would be harmed even more by Option 3 in relation to either Option 1 or Option 2.

5.2 Minimize Unanswered Complaints

Thus far, we have seen that Harm Minimization, Objection Minimization, and Avoid Reasonable Objections satisfy WDAE and ILA but violate Weaker DA, and that the Maximization Theory satisfies all the dominance principles but violates WILA. This motivates the search for a harm-avoidance theory that accommodates both WILA and the dominance principles.

The second tournament theory that we shall consider, due to Podgorski (2023),

accomplishes this. According to his *Minimize Unanswered Complaints*, when an option causes existential or comparative harm to an individual, this provides grounds for a complaint on behalf of the individual against the choice of the option, where this complaint is had *relative* to some alternative that would have either made the individual better off or not harmed them.[38]

Moreover, on this theory, existential benefits to individuals provide what Podgorski calls *'answers'* to complaints. When comparing only two options, A and B, the strength of an individual's complaint against A relative to B is the magnitude of the existential or comparative harm she incurs in the outcome of A. And if an individual exists conditional on A but not on B, then she generates an answer to harm-based complaints resulting from the choice of A *iff* her well-being conditional on A is positive, and the strength of this answer is the magnitude of her positive well-being.

When comparing only two options, A and B, if A harms some individual, then her complaint against A relative to B is unanswered, either entirely or partially, if the total of existential benefits brought about by A relative to B is less than the harm to this individual. And if there are unanswered complaints against A, then the total strength of the unanswered complaints against A relative to B is equal to the total harm of A relative to B minus the existential benefits of A relative to B.

Podgorksi's theory includes both a criterion for when one option 'defeats' another in a pairwise comparison, and a general condition of permissibility, based on the criterion of defeat. He states his criterion of defeat as follows:

> **Minimize Aggregate Unanswered Complaints*:** An option X defeats option Y *iff* the strength of unanswered complaints against X relative to Y is less than the strength of unanswered complaints against Y relative to X.[39]

The general criterion of permissibility based on the above criterion of defeat is what Podgorski calls

> **Uncovered:** An option is permissible iff there is no option that covers it, where A *covers* B *iff* A defeats B and any option(s) that B defeats.

---

[38] 'Minimize Unanswered Complaints' is our label, not Podgorski's.

[39] Notice that because a harm-based complaint is just as strong as the magnitude of the harm imposed on the complainant, and because the strength of an existential benefit answer is just as strong as the magnitude of that benefit, Podgorski's criterion of defeat can also be stated more simply in terms of harm and existential benefit, using Thomas's formalism:

> **Minimize Aggregate Unanswered Harm (MAUH):** In a binary choice, A *defeats* B iff both (i) Harm(B) − ExBen(B) > 0 and (ii) Harm(B) − ExBen(B) > Harm(A) − ExBen(A).
>
> A defeats B just in case B has at least some unanswered harm and the total unanswered harm of B is greater than that of A.

Minimize Unanswered Complaints is the conjunction of Minimize Aggregate Unanswered Complaints* and Uncovered.

Like The Maximization Theory, Podgorski's theory satisfies both WDA and Weaker DA. According to Uncovered, in any choice situation, B is permissible *iff* there is no option that covers B. But given Podgorksi's criterion of defeat, one can prove that in any choice context where A and B are both options, if A weakly addition-dominates B, then A covers B, i.e., for any C, if B defeats C, A defeats C. We prove this in the appendix.[40]

Minimize Unanswered Complaints also satisfies ILA, as well as WILA. This is because, as Podgorski points out, an important property of Uncovered is that "losers cannot dislodge winners"; if some option A is permissible, then the addition of option B can make A impermissible only if B is permissible. A theory violates ILA and WILA only when it implies that the introduction of an impermissible option can flip the deontic status of one of the other options from permissible to impermissible. But if this cannot happen, then ILA and WILA are guaranteed.

Since Minimize Unanswered Complaints satisfies WDA, Weaker DA, and WILA, it may seem like a promising harm-avoidance theory.

However, Thornley (2023) has recently raised a serious objection to Minimize Unanswered Complaints. See Table 7.[41]

Table 7. Thornley's Case

	Huckleberry	Yogi
Option 1	100	Ω
Option 2	0	2
Option 3	Ω	1

In a binary choice between Options 1 and 2, it is clear that Huckleberry has the strongest unanswered complaint against Option 2, and that therefore Option 1 defeats Option 2 according to Minimize Unanswered Complaints. However, simply introducing the possibility of making Yogi's life worse (Option 3) makes Option 2 permissible, as now there is no option that covers Option 2. In other words, there is no option that defeats Option 2 *and* any option that Option 2 defeats. For according to Minimize Unanswered Complaints, although Option 1 defeats Option 2, it does not defeat Option 3, since no one is harmed by Option 3 relative to Option 1, or by Option 1 relative to Option 3. Even worse, as Thornley (2023) notes, Option 2 will be permissible no matter how strong Huckleberry's harm-based complaint is against

[40] In the proof, the harm-based formulation MAUH is used (cf. footnote 41).

[41] Our presentation of his case was sourced, with minor cosmetic changes, from Thornley (2023, p. 524).

Option 2 relative to Option 1, and no matter how little harm Option 2 prevents from befalling Yogi relative to Option 3. Thornley calls this *'the Problem of Impairable Life Acceptance'*.

Thornley's Case demonstrates that Minimize Unanswered Complaints violates NED, a result which Podgorski, like Horton, is happy to accept. Option 3 creates a non-identity shortfall relative to Option 1, since Option 1 gives Huckleberry a life that is much better than the life that Option 3 gives Yogi; yet Minimize Unanswered Complaints entails that neither option defeats the other. One might therefore wonder whether *the Problem of Impairable Life Acceptance* could be avoided by modifying Podgorksi's criterion of defeat to reflect the apparent moral significance of non-identity shortfall as well as that of harm. Such a modified criterion would of course need to be worked out, and one would need to decide how non-identity shortfall is to be weighed against unanswered harm for the purpose of determining defeat. But the criterion would at least generate plausible results in *Thornley's Case*. It would imply that Option 1 defeats Option 3 because of Option 3's non-identity shortfall relative to Option 1, that Option 1 defeats Option 2 because of Huckleberry's harm-based complaint against Option 2 relative to Option 1, and that Option 2 defeats Option 3 because of Yogi's harm-based complaint Against Option 3 relative to Option 1. Option 1 would then cover both Option 2 and Option 3, and so Options 2 and 3 would be impermissible and Option 1 morally required, which is intuitively the correct result.

However, even such a revamped version of Minimize Unanswered Complaints would have the troubling feature that an option against which some person has an unanswered complaint, no matter how strong, can be permitted. Let us say that an amount of harm is unanswered *iff*, corresponding to that harm, there are unanswered complaints of a certain strength. Then Minimize Unanswered Complaints will sometimes permit any amount of unanswered harm, even if it is modified to account for the significance of non-identity shortfall.

For instance, consider Table 8, which represents a range of different possible addition (as opposed to non-identity) cases where x and y represent different well-being values that could obtain for Barney and Betty in these addition cases.

Table 8. Unlimited Harm

	Barney	Betty
Option 1	0	Ω
Option 2	$x+1$	$-x$
Option 3	0	y

For any $x, y > 0$, Option 2 is permissible according to Minimize Unanswered Complaints, regardless of whether its criteria of defeat imply that one option can defeat another when the latter causes non-identity shortfall relative to the former. In *Unlimited Harm*, none of the options causes non-identity shortfall. And since Barney exists given any option (i.e., he is not a contingent person), the harm to Betty done by Option 2 is wholly unanswered. Yet, no matter how awful Betty's life given Option 2 (i.e., no matter what negative value we assign to $-x$), and no matter how fabulous her life given Option 3 (i.e., no matter what positive value we assign to y), Option 2 remains uncovered, and therefore permissible. Neither Option 1 nor Option 3 can cover Option 2 on Minimize Unanswered Complaints. For any $x \geq 0$, Option 2 defeats Option 1, since the comparative harm that Option 1 does to Barney is greater than the existential harm that Option 2 does to Betty. For sufficiently large values of y in relation to x, Option 3 defeats Option 2. However, there are no values of x and $y \geq 0$, for which Option 3 defeats Option 1, since neither Option 1 nor Option 3 causes any harm or non-identity shortfall relative to the other.

Since x and y can take any values greater than 0, Option 2 can inflict any *greater* amount of permissible unanswered harm than either Option 1 or Option 3. Podgorski's Minimize Unanswered Complaints therefore violates the following principle, regardless of whether it is modified to accommodate NED:

> **Limit Permissible Harm (LPH):** If option A does more unanswered harm than any alternative, and no alternative causes non-identity shortfall, then if A is permissible, the difference between the amount of unanswered harm done by A and that done by any alternative cannot be arbitrarily great.

In other words, there must be a limit to *how much more* unanswered harm a permissible option does relative to the alternatives.

Not only is LPH intuitively plausible, it is difficult to see how any harm-avoidance theory can reject it. According to the harm-avoidance account, the only possible wrong-makers for any of Options 1–3 in *Unlimited Harm* is the harm it does, since none of the options causes non-identity shortfall. How, then, can there be *no limit* to the amount of unanswered harm that is permitted? Podgorski's proposed criterion of permissibility, Uncovered, does not track what is morally relevant on the harm-avoidance account, namely *harm-avoidance*.

6. The End of the Road

So far, we have seen that three harm-avoidance theories violate Weaker DA, a fourth satisfies Weaker DA but violates WILA, while a fifth satisfies both Weaker DA and

WILA but violates LPH. This motivates the search for a harm-avoidance theory that accommodates all the aforementioned principles.

But we've come to the end of the road. No harm-avoidance theory can accommodate all three principles. Given two very weak assumptions, which we state below, the conjunction of these principles is incompatible with the defining feature of a harm-avoidance theory. Recall that according to Harmless Permission, an option that causes no existential harm, comparative harm, or non-identity shortfall is permissible. As we now demonstrate, Harmless Permission is incompatible with the conjunction of Weaker Dominance Addition, Weak Improvable Life Acceptance, and Limit Permissible Harm.

Consider Table 9.

Table 9. The End of the Road

	Person 1	Person 2
Option 1	$-x$	Ω
Option 2	y	y
Option 3	$-x-\varepsilon$	z

The End of the Road is an abstract schema for a range of possible cases where $x, y, z,$ and ε are well-being values for Persons 1 and 2, and these values can differ across different possible cases in the range.

The schema has five important features:

Feature 1: For any $x, y, z, \varepsilon > 0$, in a choice between Option 1 and Option 3, Option 1 causes no existential or comparative harm.

Feature 2: None of Options 1—3 causes non-identity shortfall, either in a binary choice or in a choice between all three Options.

Feature 3: For any $x, y, z, \varepsilon > 0$,

> 3a. In a choice between Option 1, Option 2, and Option 3, Person 1 is the only person harmed by Option 1.
>
> 3b. Option 3 harms Person 1 more than Option 1 does.

Feature 4: For any $x, y, z, \varepsilon > 0$, and $z - y > y - (x - \varepsilon)$, Option 3 does more unanswered harm than Option 1 or Option 2.

Feature 5: For any $x, y > 0$,

5a. Option 2 weakly addition-dominates Option 1.

5b. Everyone who exists given Option 1 has a bad life.

We adopt the following definitions:

Definition 1: Option A is morally required $= df$. A is permissible and any alternative to A is impermissible.

Definition 2: Option A is impermissible $= df$. A is not permissible.

Finally, we make the following two substantive but very weak assumptions:

Weak No Dilemma Assumption: For any $x, y, z, \varepsilon, > 0$, at least one of Options 1—3 is permissible.

Weak Completeness Assumption: For any $x, y, z, \varepsilon, > 0$, and for each one of Options 1—3, that option is either permissible or impermissible.

According to Weak No Dilemma Assumption, the cases that fit the schema *The End of the Road* where $x, y, z, \varepsilon > 0$ are not moral dilemmas. At least one of the options in those cases is permissible. This does not imply that *there are no moral dilemmas*. Hence, those who believe in the existence of moral dilemmas can accept Weak No Dilemma Assumption. But we think that *if* there are any moral dilemmas, there must be a special explanation as to why, in those choice contexts, every one of an agent's options is impermissible. We do not think that there is any such explanation to be given regarding the relevant cases that fit *The End of the Road*. The onus is on those who disagree to show why Weak No Dilemmas Assumption should be rejected.

According to Weak Completeness Assumption, in the cases that fit the schema *The End of the Road* where $x, y, z, \varepsilon > 0$, for each one of Options 1—3, its deontic status is either *permissible* or *impermissible*, not some third status, such as *indeterminate*. Again, we are not assuming that the deontic status of *any* option is either permissible or impermissible, only that this is true in the relevant range of cases. Like moral dilemmas, deontic indeterminacy is a phenomenon that requires special explanation, and we just don't see what the explanation could be in the cases that fit *The End of the Road*.

Given Weak No Dilemma Assumption, Weak Completeness Assumption, and Definitions 1 and 2, we can demonstrate that Harmless Permission, Weaker Domi-

nance Addition, Weak Improvable Life Acceptance, and Limit Permissible Harm are jointly incompatible.

Proof. Assume for reductio

> **P1. Harmless Permission:** If option A does no comparative or existential harm, and does not create any non-identity shortfall, then A is permissible.
>
> **P2. Weak Improvable Life Acceptance:** If (i) option A imposes greater harm on person S than option B, and (ii) the only moral consideration against B, in a choice from an option set $\mathcal{O}$ that includes A and B, is that B harms S, then if A is not morally required in a binary choice between A and B, then A is not morally required in a choice from $\mathcal{O}$.
>
> **P3. Weaker Dominance Addition:** If option A weakly addition-dominates option B, and everyone who exists given B has a bad life, then B is impermissible.
>
> **P4. Limit Permissible Harm:** If option A does more unanswered harm than any alternative, and no alternative causes non-identity shortfall, then if A is permissible, the difference between the amount of unanswered harm done by A and that done by any alternative cannot be arbitrarily great.

From P1, Feature 1, and Definitions 1 and 2,

> **P5.** For any $x, y, z, \varepsilon > 0$, in a binary choice between Option 1 and Option 3, Option 3 is not morally required.

From P1, P2, P5, and Features 2 and 3,

> **P6.** For any $x, y, z, \varepsilon > 0$, in a choice between Option 1, Option 2, and Option 3, Option 3 is not morally required.[42]

From P6, Weak No Dilemma Assumption, Weak Completeness Assumption, and Definitions 1 and 2,

[42] Notice that P1 and Features 2 and 3 jointly imply that if Option 1 is impermissible in a choice between Options 1—3, then this can only be because Option 1 harms Person 1. In other words, the fact that Option 1 harms Person 1 is the only moral consideration against Option 1.

> **P7.** For any $x, y, z, \varepsilon > 0$, and $z - y > y - (-x - \varepsilon)$, in a choice between Option 1, Option 2, and Option 3, either Option 1 is permissible or Option 2 is permissible.

From P4, and Features 2 and 4,

> **P8.** For some $x, y, z, \varepsilon, > 0$, and $z - y > y - (-x - \varepsilon)$, in a choice between Option 1, Option 2, and Option 3, Option 2 is impermissible.

From P7, P8, and Definition 2,

> **P9.** For some $x, y, z, \varepsilon, > 0$, and $z - y > y - (-x - \varepsilon)$, in a choice between Option 1, Option 2, and Option 3, Option 1 is permissible.

But from P3 and Feature 5,

> **P10.** For any $x, y, z, \varepsilon, > 0$, in a choice between Option 1, Option 2, and Option 3, Option 1 is impermissible.

So, from P10, Definition 2, and existential instantiation,

> **C.** It is not the case that for some $x, y, z, \varepsilon, > 0$, and $z - y > y - (-x - \varepsilon)$, in a choice between Option 1, Option 2, and Option 3, Option 1 is permissible.

Since C contradicts P9, we must reject either Harmless Permission, Weaker Dominance Addition, Weak Improvable Life Acceptance, or Limit Permissible Harm.

But Harmless Permission is part and parcel of any harm-avoidance theory. So it seems, proponents of a harm-avoidance theory must reject either Weaker Dominance Addition, Weak Improvable Life Acceptance, or Limit Permissible Harm. The worry is that each of these claims is extremely plausible, more plausible, we think, than Harmless Permission.

7. Conclusion

One of the central challenges facing any harm-avoidance theory, i.e., any theory committed to Harmless Permission, is offering an adequate response to *the Problem of Improvable Life Avoidance*. The problem is that the simplest harm-avoidance theory, Harm Minimization, leads to both improvable life avoidance and a requirement to choose weakly addition-dominated options. Most harm-avoidance theorists seem to

agree that improvable life avoidance and requiring weakly addition-dominated options are problematic. But in scrutinizing the existing harm-avoidance theories, and their responses to *the Problem of Improvable Life Avoidance*, we have argued that some of these theories permit weakly addition-dominated options in which everyone has a bad life, that some lead to strong improvable life avoidance, and that some permit any amount of harm in relation to less harmful alternatives. Moreover, we have argued that no harm-avoidance theory can avoid all three of these problems.

Our discussion bears on the prospects of finding a general theory that accommodates the Asymmetry. Harm-avoidance theories have seemed like the most promising candidates in this regard. In light of our discussion, one might be motivated to find an alternative theoretical framework in which to situate the Asymmetry. But the alternatives come with their own problems.

One possibility would be to defend the Asymmetry by appealing to a different type of harm-avoidance. For instance, one could claim that the only harm that we are morally required to avoid is non-comparative harm. We might be required to avoid causing people to be in an intrinsically bad state, but not to avoid giving people less of what is intrinsically good rather than more of what is intrinsically good.

However, this may seem quite extreme. It implies, for example, that we have no moral requirement to save people from death, insofar as death would not be intrinsically bad for those who die but would merely deprive them of further good.

An alternative response, which we find more plausible, is to reject the Asymmetry. We should accept that we *can* be morally required to create people with good lives rather than not create them at all, where the explanation for this is simply that these people would exist with good lives.

What about Climate Anti-Natalism, the claim that in many situations it is wrong to create a person because of the added CO2 emissions? Insofar as Climate Anti-Natalism is motivated by the harm-avoidance account, our result undermines the case for Climate Anti-Natalism. In the domain of normative population ethics, there are several sources of an act being impermissible. It could be impermissible because it causes harm, because it causes non-identity shortfall, or because it fails to create people with good lives. We have argued against the claim that an act that causes no harm or non-identity shortfall in this domain is permissible. This makes it seem likely that there will be cases where an act is impermissible because it fails to create a person with a good life. Those who wish to defend Climate Anti-Natalism must therefore address the possible existence of such reasons, and show that they aren't strong enough to outweigh the expected climate-change-related harm of adding another person to the world.

References

Arrhenius, G. (2003). The person-affecting restriction, comparativism, and the moral status of potential people. *Ethical Perspectives*, *10(3)*, 185–195.

Arrhenius, G. (2022). Population paradoxes without transitivity. In G. Arrhenius Arrhenius, K. Bykvist, T. Campbell, & E. Finneron-Burns (Eds.), *The Oxford handbook of population ethics* (pp. 181–203). Oxford University Press.

Arrhenius, G., Bykvist, K., Campbell, T., & Finneron-Burns, E. (2022). Introduction. In G. Arrhenius, K. Bykvist, T. Campbell & E. Finneron-Burns (Eds.), *The Oxford handbook of population ethics* (pp. 1–12). Oxford University Press.

Bader, R. (2022a). The asymmetry. In J. McMahan, T. Campbell, J. Gooddrich & K. Ramakrishnan (Eds.), *Ethics and existence: The legacy of Derek Parfit* (pp. 15– 37). Oxford University Press.

Bader, R. (2022b). Person-affecting utilitarianism. In G. Arrhenius, K. Bykvist, T. Campbell & E. Finneron-Burns (Eds.), *The Oxford handbook of population ethics* (pp. 251–270). Oxford University Press.

Benatar, D. (2006). *Better never to have been: The harm of coming into existence*. Oxford University Press.

Boonin, D. (1996). Don't stop thinking about tomorrow: Two paradoxes about duties to future generations. *Philosophy & Public Affairs*, *25(4)*, 267–307.

Boonin, D. (2014). *The non-identity problem & the ethics of future people*. Oxford University Press.

Bradley, B. (2012). Doing away with harm. *Philosophy and Phenomenological Research*, *85(2)*, 390–412.

Burkett, Daniel. (2021). A Legacy of Harm? Climate Change and the Carbon Cost of Procreation. *Journal of Applied Philosophy* 38: 790-808.

Bykvist, K. (2006). The benefits of coming into existence. *Philosophical Studies*, *135(3)*, 336–362.

Chappell, R. Y. (2017). Rethinking the asymmetry. *Canadian Journal of Philosophy*, *47(2-3)*, 167–177.

Conly, S. (2015). One Child: Is there a right to more? New York: Oxford University Press.

Cusbert, J., & Kath, R. (2018). A consequentialist account of Narveson's dictum. *Philosophical Studies*, *176(7)*, 1693–1709.

Frick, J. (2020). Conditional reasons and the procreation asymmetry. *Philosophical Perspectives*, *34(1)*, 53–87.

Frick, J. (2022). Context-dependent betterness and the mere addition paradox. In J. McMahan, T. Campbell, J. Gooddrich & K. Ramakrishnan (Eds.), *Ethics and existence: The legacy of Derek Parfit* (pp. 232–263). Oxford University Press.

Gardner, M. (2015). A harm-based solution to the non-identity problem. *Ergo*, *2(17)*, 427–444.

Gardner, M. (2019). David Boonin on the non-identity argument: Rejecting the second premise. *Law, Ethics and Philosophy*, *7*, 29–47.

Gert, J. (2004). *Brute rationality*. Cambridge University Press.

Harman, E. (2009). Harming as causing harm. In M. Roberts & D. Wasserman (Eds.), *Harming future persons: Ethics, genetics and the nonidentity problem* (pp. 137–154). Springer.

Hedberg, T. (2019). "The duty to reduce greenhouse gas emissions and limits of permissible procreation. *Essays in Philosophy* 20: 1–24.

Herlitz, A. (2020). Non-transitive better than relations and rational choice. *Philosophia*, *48(1)*, 179–189.

Holtug, N. (2010). *Persons, interests, and justice*. Oxford University Press.

Horton, J. (2021). New and improvable lives. *Journal of Philosophy*, *118(9)*, 486–503.

Kaczmarek, P. & Lloyd, H. R. (forthcoming). Moral uncertainty, pure justifiers, and agent-centred options. *Australasian Journal of Philosophy*, *00*, 1–29. https : //philpapers.org/rec/KACMUP-2.

Kagan, S. (1991). Replies to my critics. *Philosophy and Phenomenological Research*, *51(4)*, 919–928.

Kamm, F. M. (1985). Supererogation and obligation. *Journal of Philosophy*, *82(3)*, 118– 138.

Kamm, F. M. (1996). *Morality, mortality, vol. 2, rights, duties, and status*. Oxford University Press.

Lazar, S. (2013). Associative duties and the ethics of killing in war. *Journal of Practical Ethics*, *1(1)*, 3–48.

Little, M. O., & Macnamara, C. (2021). Non-requiring reasons. In R. Chang & K. Sylvan (Eds.), *The Routledge handbook of practical reasons* (pp. 393–404). Routledge.

MacIver, C. (2015). Procreation or appropriation? In Sarah Hannan, Samantha Brennan & Richard Vernon, eds. *Permissible Progeny?: The Morality of Procreating and Parenting*, 107-128. Oxford University Press.

McDermott, M. (1982). Utility and population. *Philosophical Studies*, *42(2)*, 163–177.

McDermott, M. (2019). Harms and objections. *Analysis*, *79(3)*, 436–448.

McMahan, J. (1981). Problems of population theory. *Ethics*, *92(1)*, 96–127.

McMahan, J. (2013). Causing people to exist and saving lives. *Journal of Ethics*, *17(1/2)*, 5–35.

Meacham, C. J. (2012). Person-affecting views and saturating counterpart relations. *Philosophical Studies*, *158(2)*, 257–287.

Mogensen, A. (2019). Staking our future: Deontic long-termism and the non-identity problem. *Global Priorities Institute Working Paper No. 9-2019*, *00*, 1–32. https: //globalprioritiesinstitute.org/andreas- mogensen- staking- our- futuredeontic-long-termism-and-the-non-identity-problem/.

Mogensen, A. (2021). Moral demands and the far future. *Philosophy and Phenomenological Research*, *103(3)*, 567–585.

Muñoz, D. (2021). From rights to prerogatives. *Philosophy and Phenomenological Research*, *102(3)*, 608–623.

Otsuka, M. (1997). Kamm on the morality of killing. *Ethics*, *108(1)*, 197–207.

Otsuka, M. (2017). How it makes a moral difference that one is worse off than one could have been. *Politics, Philosophy & Economics*, *17(2)*, 192–215.

Parfit, D. (1984). Reasons and Persons. Oxford: Oxford University Press.

Parfit, D. (2017). Future people, the non-identity problem, and person-affecting principles. *Philosophy & Public Affairs*, *45(2)*, 118–157.

Persson, I. (2017). *Inclusive ethics: Extending beneficence and egalitarian justice*. Oxford University Press.

Podgorski, A. (2023). Complaints and tournament population ethics. *Philosophy and Phenomenological Research*, *106(2)*, 344–367.

Pummer, T. (2023). *The rules of rescue: Cost, distance, and effective altruism*. Oxford University Press.

Roberts, M. (2007). The non-identity fallacy: Harm, probability and another look at Parfit's depletion example. *Utilitas*, *19(3)*, 267–311.

Roberts, M. (2011). The asymmetry: A solution. *Theoria*, *77(4)*, 333–367.

Ross, J. (2015). Rethinking the person-affecting principle. *Journal of Moral Philosophy*, *12(4)*, 428–461.

Schwartz, T. (1972). Rationality and the myth of the maximum. *Noûs*, *6(2)*, 97–117.

Shiffrin, S. (1999). Wrongful life, procreative responsibility, and the significance of harm. *Legal Theory*, *5*, 117–148.

Spencer, J. (2021). The procreative asymmetry and the impossibility of elusive permission. *Philosophical Studies*, *178(11)*, 3819–3842.

Temkin, L. (2012). *Rethinking the good*. Oxford University Press.

Thomas, T. (2022). The asymmetry, uncertainty, and the long term. *Philosophy and Phenomenological Research*, *107(2)*, 470–500.

Thornley, E. (2023). The procreation asymmetry, improvable-life avoidance and impairablelife acceptance. *Analysis*, *83(3)*, 517–526.

Wynes, S. and Nicholas, K. (2017). The climate mitigation gap: Education and government recommendations miss the most effective individual actions. *Environmental Research Letters* 12: 3.

Young, T. (2001). Overconsumption and procreation: Are they morally equivalent? *Journal of Applied Philosophy* 18 (2001): 183–192.

Appendix. Minimize Unanswered Complaints Satisfies Weak Dominance Addition

We will show that on Podgorski's Minimize Unanswered Complaints:

> **Claim:** For any choice context in which A and B are both options, if A weakly addition dominates B, then A defeats B and for any C, if B defeats C, then A defeats C.

It follows from Claim that if A weakly addition dominates B, then A covers B, and thus B is impermissible on the Uncovered criterion. So, the conditional 'if B is permissible, then A is permissible' is vacuously true.

> *Proof.* Suppose that in some choice context,
> **P1.** A weakly addition dominates B.

First we will prove that given P1, A defeats B. From P1 and the definition of 'weak addition dominance',

> **P2.** Every person who exists in B exists in A, and every person who exists in B has well-being at least 0 and at most x, and every person who exists in A has positive well-being $y > x$.

Let $\text{Harm}(A_B)$ represent the total harm in A relative to B. Because total harm is the sum of existential harm and comparative harm, from P2 and the definitions of 'existential harm', 'comparative harm', and 'existential benefit answers' we derive P3—P5:

> **P3.** $\text{Harm}(A_B) = 0$.
>
> **P4.** $\text{Harm}(B_A) > 0$.
>
> **P5.** $\text{ExBen}(B_A) = 0$.
>
> From P4 and P5, we derive
>
> **P6.** $\text{Harm}(B_A) - \text{ExBen}(B_A) > 0$

From P3 and P6, we derive

> **P7.** $\text{Harm}(B_A) - \text{ExBen}(B_A) > \text{Harm}(A_B) - \text{ExBen}(A_B)$.

From P6, P7, and Minimize Aggregate Unanswered Harm's definition of 'defeat', i.e., MAUH, we derive

P8. A defeats B.

Next, we will prove that for any alternative C in the choice context, if B defeats C, then A defeats C. Suppose:

P9. There is some C in the choice context, such that B defeats C.

Since B defeats C, by the definition of 'defeat',

P10. $\text{Harm}(C_B) - \text{ExBen}(C_B) > 0$

and

P11. $\text{Harm}(C_B) - \text{ExBen}(C_B) > \text{Harm}(B_C) - \text{ExBen}(B_C)$.

From P10, we derive

P12. $\text{Harm}(C_B) > 0$.

We will show that for any individual harm in C relative to B, there is at least that much individual harm in C relative to A.

First, any harm in C relative to B is either existential harm in C relative to B or comparative harm in C relative to B.

Any existential harm in C relative to B is suffered either by someone who exists in C but neither A nor B, or by someone who exists in C and A but not in B. Any person who suffers existential harm in C relative to B, has negative well-being $-z$ in C. If the person exists in C but neither A nor B, then the magnitude of her existential harm in C relative to B and that of her existential harm in C relative to A is $|-z|$. If she exists in C and A but not B, then the harm she suffers in C relative to A is comparative, and the magnitude of this harm in C relative to A is $|-z|+y$, i.e. the difference between her positive welfare in A and her welfare in C, where $(|-z| + y) > |-z|$. Hence, for any existential harm in C relative to B, there is either that much existential harm in C relative to A, or even greater comparative harm in C relative to A.

Next, any comparative harm in C relative to B is suffered by someone who exists in both B and C. Suppose she has lifetime well-being w_C in C, where w_C could be any positive or negative number, $w_C < x$. Since everyone who exists in B exists in A with welfare y, and $y > x$, it straightforwardly follows that $w_C < x < y$. Hence, any

comparative harm in C relative to B is an even greater comparative harm in C relative to A.

It follows that for any individual harm in C relative to B, there is at least that much individual harm in C relative to A. Since the total harm in one option relative to another is just the sum of the individual harms in the former relative to the latter, we derive

P13. $\text{Harm}(C_A) \geq \text{Harm}(C_B)$.

From P10 and P13, we derive

P14. $\text{ExBen}(C_B) < \text{Harm}(C_A)$.

By the definition of 'existential benefits', any existential benefit in C relative to B is had by a person who exists in C but not B. Hence, any existential benefit in C relative to B is had either by a person who exists in C but not in A or B, or by a person who exists in C and A, but not B. Hence,

P15. $\text{ExBen}(C_B)$ = the sum of positive well-being of (i) all people who exist in C but not in A or B and (ii) all people who exist in C and A but not B.

Let W_C = the sum of positive well-being of all people who exist in C but not A or B. And let W_{CA} = the sum of positive well-being of all people who exist in C and A but not B.

Then, from P14 and P15, we derive

P16. $W_C + W_{CA} < \text{Harm}(C_A)$.

From P16, we derive

P17. $W_{CA} < \text{Harm arm}(C_A)$ and

P18. $W_C < \text{Harm}(C_A)$.

The existential benefits of C relative to A consists of the positive well-being of those who exist in C but not A. Since everyone who exists in B exists in A, this means that the existential benefits of C relative to A consists of the positive well-being of those who exist in C but neither A nor B, i.e.,

P19. $\text{ExBen}(C_A) = W_C$.

From P18 and P19, we derive

P20. $\text{ExBen}(C_A) < \text{Harm}(C_A)$.

Subtracting $\text{ExBen}(C_A)$ from both sides of the inequality in P20, we derive

P21. $\text{Harm}(C_A) - \text{ExBen}(C_A) > 0$

Next, we need to see how $\text{Harm}(A_C)$ compares to $\text{Harm}(C_A)$.

There are no existential harms in A relative to any outcome. Hence, if there is any individual harm in A relative to C, it is a comparative harm, i.e., a harm to someone who exists in A and C. Everyone who exists in A has positive well-being *y*. By the definition of 'comparative harm', for anyone harmed in A relative to C, her well-being in C is greater than her well-being in A, (i.e., $> y$). Call those harmed in A relative to C '*the A-harmed people*'. Let $W_{\text{A-harmed}}{}^{A}$ be the sum total of positive well-being of the A-harmed people in A. Let $W_{\text{A-harmed}}{}^{C}$ be the sum total of positive well-being of the A-harmed people in C. The total harm of A relative to C is therefore equal to $W_{\text{A-harmed}}{}^{C} - W_{\text{A-harmed}}{}^{A}$. In other words,

P22. $\text{Harm}(AC) = W_{\text{A-harmed}}{}^{C} - W_{\text{A-harmed}}{}^{A}$.

Since W_{CA} is the total positive well-being of everyone who exists in both A and C,

P23. $W_{\text{A-harmed}C} \leq W_{CA}$.

Since $W_{\text{A-harmed}}{}^{A}$ is a positive number, from P23, we derive

P24. $W_{\text{A-harmed}C} - W_{\text{A-harmed}A} < W_{CA}$.

From P22 and P24, we derive

P25. $\text{Harm}(A_C) < W_{CA}$.

From P16, P19, and P25, we derive

P26. $\text{ExBen}(C_A) + \text{Harm}(A_C) < \text{Harm}(C_A)$.

Subtracting $\text{ExBen}(C_A)$ from both sides of the inequality in P26, we get:

P27. $\text{Harm}(A_C) < \text{Harm}(C_A) - \text{ExBen}(C_A)$.

The unanswered harm of C relative to A is greater than the harm of A relative to C, and hence, greater than the unanswered harm of A relative to C. Finally, from P21, P27, and the definition of 'defeat', **C.** A defeats C

We have proven that if A weakly addition dominates B, then A defeats B and for any C, if B defeats C, A defeats C. From which it follows that if A weakly addition dominates B, A covers B. Therefore, Podgorski's Minimize Unanswered Complaints satisfies Weak Dominance Addition.

Säde Hormio[1]

Droplets of Detriment and Pint-Sized Profits: Small Contributions to Collective Outcomes[2]

Moral theories struggle to give a reason why individuals should or should not contribute to a collective outcome when the contribution is small enough to make no relevant difference to it. This is problematic if most contributions that make up a normatively important outcome share this feature. Although the literature on the problem of small contributions has focused on momentary token choice situations, I will argue that the central question should instead be individual behaviour over time and contributions to certain types of outcomes. Because most real-life cases are about collective outcomes that aggregate over time, the crucial question is not about contributions to a harm (or failing to help) on some specific one-off occasion. Instead, what matters more is if we regularly perform, or try to avoid, that type of contribution. I argue that in many cases, the correct unit of moral analysis is not the individual act, but the coherence of the moral life of a person. Failing to act according to our individual values in collective settings compromises our integrity as moral agents. If one attempts to separate the individual and the collective domains starkly in moral matters, it can lead to a lack of coherence between one's values and contributions.

[1] Practical Philosophy, University of Helsinki & Institute for Futures Studies, Stockholm (sade.hormio@helsinki.fi)

[2] Financial support from Riksbankens Jubileumsfond (grant no. P22-0662) is gratefully acknowledged.

1. Introduction

Individual contributions to collectively caused outcomes can sometimes be very small, even so tiny as to make no relevant difference. Yet, in aggregate, these small contributions can result in a morally significant outcome, whether good or bad. Moral theories struggle to explain why individuals should make or withhold such contributions, which is problematic, as many global ills today are best described as small contributions to a great harm (Kutz 2000; Lichtenberg 2014; Nefsky 2019). Some prominent examples discussed in the literature include microplastics found in the oceans, greenhouse gases emitted around the world accumulating in the atmosphere to cause climate change, or customers buying products made with sweatshop labour. It seems that we want to give individuals a reason not to make these contributions, because if no contributor feels that they should act differently, there is a danger that tackling many modern harms will appear to be the responsibility of no one. Of course, there are usually collective agents who bear responsibility: the manufacturer of plastic bottles, the retailer of overly cheap clothing, or the fossil fuel industry digging up more coal, oil, and gas. There are also collective agents that potentially have extensive power to change practices, including nation-states, international regulators, or investors. Even so, it is not clear that so-called 'responsibility gaps' would not appear: the actions of collective agents might not account for the entirety of the harm, or the collective agents might not have the tools and power to fix the whole problem by themselves (Collins 2019).[3]

When a collectively caused outcome is morally significant, it should give individuals a reason to contribute to it, or to refrain from doing so, but if the contributions make no difference, it seems hard to pinpoint what that normative reason is.[4] I will call this the *Problem of Small Contributions*. The problem can be discussed in terms of what moral reasons are there for (not) performing acts of a certain type (such as buying clothes made in sweatshops), or what moral reasons are there for (not) performing some particular token act (like buying a dress from an online retailer known for using sweatshops during their promotional campaign to get new customers). I will label these *Type Problem of Small Contributions*, and *Token Problem of Small Contributions*, respectively. The former is easier to solve.

There is an ongoing debate among philosophers about the normative significance

[3] There might also be motivational gaps in getting collective agents to change their course of action without individuals insisting that they do so. There are also cases in which the harm is caused by widespread harmful practices and structures that are outside the full control of even powerful collective agents (think of ingrained racism in a society).

[4] I assume, uncontroversially, that people are motivated to do something that they judge to be the right thing, without taking sides in the debates over the exact link between moral motivation and judgement, or on the strength of the motivation.

of very small contributions (e.g. Asker 2023; Barnett 2018; Broome 2019; Budolfson 2019; Kagan 2011; Kutz 2000; Nefsky 2017; Spiekermann 2014). Some argue that all small contributions to great harms are also harmful by themselves. Others disagree and argue that small contributions become harms only when combined with enough other such contributions, a position that could be labelled as *harmless in isolation*. Although not taking sides in that debate, I will argue that we should care about individual contributions even if they are harmless in isolation and make no difference to an outcome by themselves. There might not always be a reason for (not) performing a particular act on a given occasion (i.e. the *Token Problem* persists), but we can have a moral *pro tanto* reason we should or should not perform certain types of acts (offering a solution to the *Type Problem*).[5]

Although the literature so far has focused on the *Token Problem of Small Contributions*, that is, on momentary choice situations, I will argue that more attention should be paid to the *Type Problem of Small* Contributions. The emphasis thus should be on individual behaviour over time. After all, moral reasons for individual actions not only stem from what the individual can do in terms of affecting the outcome, but also from concerns about what kind of person one should be, and how one should not associate oneself with wrongness. If you overlook the effects that the collectives you belong to have on the world, this can have a corrupting effect on your character. (Such an attitude could also have a corroding effect on communities and societies when widely shared, but here I will focus on the individual). However, the account I propose is not fixated on clean hands and perfect character. Rather, it is about not discounting the collective contexts within which we act.

I suggest that we should care about our marginal contributions as tokens of harmful or beneficial patterns, even when not every instance counts. The focus is thus on contributions to types of collective harms or benefits, rather than on a particular token. This is because most of the real-life cases that display the structure of the *Problem of Small Contributions* are about harms that aggregate over time (like greenhouse gas emissions, plastic waste accumulating in the oceans, the economic structures that make sweatshops viable, and so on), and not one-off situations. The same goes for many collective cases of benefit: they only come about if enough contribute over time, like is the case with donations to most charities, for example. For this reason, the crucial question in most real-life cases is not if we contribute to a harm or fail to help on some specific occasion (token), but if we regularly try to avoid or perform that type of contribution (type).

I begin by highlighting the key features of the Problem of Small Contributions

[5] Pro tanto reasons are important reasons that should be taken seriously, but which can be outweighed by other reasons all things considered, depending on the circumstances.

through two examples. Then I discuss the limits of solutions based solely on aggregate of individual effects in section three by distinguishing what those explanations must assume in order to work. These features are not present in all examples of small contributions to collectively caused outcomes. In section four, I argue for the importance of a coherent moral life of a person and how it is the correct unit of moral analysis instead of isolated individual acts, and hence the evaluative stance we should take. I also defend my integrity approach from objections.

2. Box of Doom and Rice Grains

This section highlights the key features of the Problem of Small Contributions to collectively caused outcomes through two examples, one towards a harm, the other towards a beneficial outcome. The latter cases are discussed less often in the literature, but I find them to be equally important, as omitting to make a small contribution to beneficial causes can translate into missed opportunities to make things better, or it can help to maintain a detrimental status quo. But let us start off with the harm or, more precisely, the *Box of Doom*.

> *Box of Doom*
> There is an island with 10,000 inhabitants and a Box of Doom. The Box was brought to the island a few hundred years ago by a mad scientist, who wired all the 1,000 wells on the island to connect to the Box. Each time someone pumps water from a well, a vent opens up that allows natural gas to flow into the Box. Nothing happens, as the container is very large, and the gas slowly evaporates from little holes at the top. The islanders know about the Box of Doom and have made estimates about the rate of evaporation and how much gas it can safely contain. However, if the islanders pump water over a certain level in any given year, gas starts accumulating in the Box. If enough gas builds up, it leads to an explosion after a few years. The results could be potentially catastrophic for the island. This is why each islander is aware of a safe amount to pump per year. The amount of water is enough to cover basic needs, although being able to pump more would certainly make life easier.

The key elements of the Problem of Small Contributions to a normatively significant outcome are included in the example. Firstly, individual acts do not cause the harmful outcome in isolation, but only in association with enough other such acts. In other words, they are harmless or non-impactful in isolation. As individual acts in isolation do not set off the explosion, they cause no harm as such, although they can increase the risk of harm. (Arguably, the psychological effect of the latter could be harmful.

Still, they do not cause material harm in isolation.[6]) Secondly, unilateral action is not enough to avoid the harm. There is no off-switch on the Box that someone could just flick and be done with the threat. Thirdly, the harm is not intended as such, but rather it is a side effect of some other activity. In this, the example is similar to many real-life environmental harms, like microplastics accumulating in the oceans. This is not a necessary feature of the Problem of Small Contributions, but it is often present.

Another thing to note about the example is that the explosion in the *Box of Doom* is a threshold harm, which can take many forms. In this case, there is no explosion if the yearly thresholds are not exceeded, because the gas steadily evaporates from the holes at the top. The threshold is thus met or unmet on a rolling basis. There can also be cases in which the small contributions steadily accumulate over time, so that the harm becomes more and more likely with each contribution, or there might be many thresholds.

Next, let us look at small contributions to a positive outcome.

Rice Grains

A village with 100 residents hosts a weekly party, to which they invite people from nearby villages. The neighbouring villages are less well-off, and the weekly gatherings help them to prevent malnutrition. The tradition is to serve risotto at the party, so a large pan is set up, in which risotto is cooked from rice donated by the residents. As decided by the village council (of which all residents are members), each resident is supposed to donate one cup of risotto rice each week, which equals roughly 5,000 grains. The risotto therefore has approximately 500,000 grains of rice in it. Experience has taught the villagers that this is a good size because all partygoers get enough to eat. This is also the surplus amount that each villager can donate without their own families going hungry.

As before, the outcome is brought about only if enough people contribute to it. Individual donations will only make the party risotto possible if there are enough other such donations. Furthermore, unilateral action is not enough: no villager has enough rice to make the risotto happen on their own. In the same way, an individual unilaterally opting out will not jeopardise the outcome: a risotto with 495,000 grains will still feed all the partygoers. Individual small contributions are non-impactful in isolation.

A difference with regard to the harm example is that the risotto is an intended outcome of the small contributions, not a side effect of some other activity. Another difference is that while the explosion was purely a threshold outcome, under or above

[6] Pumping over the limit might cause anxiety or stress among islanders who know about the activity. However, for the sake of simplicity, I will focus on material harm or a risk of such harm.

which individual acts do not have an impact, in *Rice Grains* there is a range for the ideal number of contributions, but no clear threshold for when the heap of rice becomes the party risotto. Individual acts over and above the range of the ideal number of contributions still have an impact, although the outcome might be sub-optimal (there is not enough food to feed all, or there is too much food, resulting in waste). Despite these differences, in both examples, a unilateral withdrawal of one's small contribution will not change the outcome for the better or the worse: there will still be enough risotto for everyone, even if Rosa does not donate, and there might be an explosion even if Riko never pumps water over the limit. These actions are non-impactful in isolation (i.e. if only one person acts in that way, the collective outcome does not come about), but if enough people cooperate with the safety limits, there will be no explosion, and if enough people do not contribute, there would not be enough risotto. What others do matters.

The configuration of real-life small contributions to collectively caused outcomes might of course bear little resemblance to such invented examples. There might be no thresholds at all, just a steady accumulation of harm or benefit. One's small contribution could also have an impact many times over, as with anthropogenic climate change. While there has been a lot of debate over the effects of individual emissions (e.g. Cripps 2016; Sandberg 2011; Kingston & Sinnott-Armstrong 2018), climate change is not the best example of the Problem of Small Contributions. This is because while individual emitting choices, such as Sunday joyrides with gas-guzzling cars, are small contributions to a great harm, they lack the central feature of being non-impactful in isolation. An individual's emissions have countless opportunities to cause harm over the decades and centuries that they spend in the atmosphere (Broome 2019). However, not all small emissions contributions are borne out of such direct choices as deciding to go on a Sunday joyride (more on this in the next section). Therefore, the Problem of Small Contributions returns.

3. Direct Small Effects and Direct Small Choices

Examples of small contributions usually assume what I will call a *direct small choice*: the agent has a choice that is entirely up to them, albeit without control over the collective outcome. They often also assume a *direct small effect*: the agent's action has an effect, even if this is imperceptible in isolation from other such acts. Yet these features are not always present in real-life instances of the problem of small contributions.

In Derek Parfit's (1987: 80–81) famous case of a mistake in moral mathematics, *The Harmless Torturers*, a thousand torturers each turns a dial that distributes a miniscule amount of pain to a thousand victims. Although no single torturer can be said

to have made their pain worse, each victim is in severe pain as they are being tortured by a thousand people. Each small contribution to the torture has a direct small effect, albeit imperceptible. The individual torturers cannot – in isolation – decide if the person is being tortured or not, or the level of pain that the victim is under. However, they have a direct small choice: it is up to them if they turn the dial or not. *Rice Grains* also has these features: each individual can decide to contribute their cup, or not, in isolation from the other decisions (direct small choice) and as a result of this decision, there is either one cup more or one cup less in the risotto (direct small effect). Although your contribution might not be perceptible or significant, and you have no direct control over the outcome, you still have some direct small room for manoeuvre.

While an individual turning the dial results in too small a difference in pain for the victim to notice, each torturer has turned the dial a thousand times. They should care because of the aggregate impacts can amount to a great harm. However, this approach works only in cases when the agent is repeatedly contributing to some outcome via direct small choices through actions which have direct small effects. But not all cases of small contributions to collective outcomes have such features. When the available infrastructure offers no real options, an average individual has no direct small choice (apart from trying to influence others). Consider contributions to environmental pollution. You can take the metro to work instead of driving your car, only if an efficient enough public transportation system is available. The same goes for the type of energy infrastructure that powers the public spaces you use (Hormio 2024: 7–8). In such cases, there is no direct small choice in isolation from others (or if there is, it is unfeasibly prohibitive: drive your car to work or quit). Compare this with the *Box of Doom*, in which each individual chooses if they abide by the water restrictions, even though they have no control over the collective outcome.

The causal impact of small contributions has been emphasised in the literature by looking at the aggregate impact made by each agent over a period of time, or at the aggregate impact on a beneficiary or victim. This is where the moral mathematics comes in: you should consider the collective setting. Although the electric shock caused by turning the dial is too small to notice, taken as a set, the torturers inflict great suffering on their thousand victims (Parfit 1987). There is an epistemic dimension to the argument about sets. In the *Box of Doom*, the individual rule-breakers do not know how many others are pumping over the limit. They are behaving in a way that would cause harm if enough other people behaved like them, but their actions do not cause material harm because there are only 70 of them (let us assume this is 20 islanders short of a 90-person-harm-causing set). Regardless of this, it is wrong for them to act this way, as they do not know what others are doing: the set they are part of could be large enough to cause an explosion. Individual contributors do wrong if they ignore the risk that their actions may become perceivable depending on what

others do (Spiekermann 2014: 89), and we should not ignore such risk because we cannot be sure what others will do. Or, as Julia Nefsky (2017) argues, we should not decide in advance that our individual action is insignificant when an outcome is uncertain.[7]

However, if you have perfect knowledge of a situation, an individual act that is harmless in isolation seems permissible if you are certain of what others do. It would be fine to pump over the limit if you have installed security cameras at all the pumps and know that enough others are complying with the restrictions. It is equally fine to turn the dial if you know for certain that your contribution will not be perceptible due to what others have done. It is unsatisfactory to have an account with such limitations, especially since it seems to allow for cases that go against our moral intuitions about how we should treat each other.

Indeed, there seems to be something strange about wondering if our individual small contribution is harmful when such activity in general causes harm. How many of us would be comfortable about being friends with harmless torturers? Surely there is something amiss in your character if you think that while torture is bad, it is acceptable for you to torture just a tiny bit, as you or someone else has calculated that it makes no difference to the suffering of the victim. It is like saying that although I do not believe in animal cruelty, because a kitten is already sure to drown in a bucket, I might as well add some more water into it. I find the cases with more bite to be the ones in which the activity is not arguably harmful already by definition (like torturing someone). These are cases like the *Box of Doom*, when you are pumping water to meet food and hygiene needs better, not to kill kittens.

4. Coherence of the Moral Life of a Person

Although we can focus on individual acts and analyse them, their normative significance can often only be evaluated by looking at the coherence of the moral life of a person, or so I suggest in this section. Although the problem of small contributions is usually framed in terms of how individual actions are instrumental to morally significant outcomes, I argue that what matters more is the issue of the individual's moral character and that potential solutions should focus on this instead. I explain what I understand by coherence and integrity, as well as discuss how contributing to a collective outcome should be conceptualised. I will also quickly note how my suggestion differs from the version of an integrity account that is said to suffer from the superfluity problem (Nefsky 2019; Wieland & van Oeveren 2020). But to begin, I start

[7] She distinguishes between making a difference and helping to bring something about (Nefsky 2017).

by explaining the morally evaluative stance that I think is often the most relevant when looking at collective impact cases.

We should consider whether an action in a collective context is something that we want to do, not only because of the possible impact of the *horizontal accumulation* (i.e. that it belongs to a set of acts that in aggregate cause harm or benefit), but also the impacts of *vertical accumulation on one's character*, to use my own terms. By choosing to use the words 'horizontal' and 'vertical' in this context, I want to draw attention to how these are different and independent dimensions of accumulation of small impacts into something normatively significant.

Think of each contribution as a dot. In horizontal accumulation, the dots are spread widely over many points, because the collective outcome is an aggregate of contributions by several people. This is how small contributions are usually described in the literature. In contrast, I use the term 'vertical' to highlight how the accumulation can also be framed from the point of view of one person (dots piled on top of one another). But instead of focusing on the aggregate causal impact of such vertical accumulation of one person's choices over a period of time (e.g. Broome 2012; Nolt 2011), I want to focus on the impact of the accumulation on the people themselves.

My argument is that such a vertical evaluative stance, focused on the coherence of the moral life of a person, can give us a reason to make or refrain from making small contributions. Reasons of character can make us rethink our contribution to a collective outcome, even if we are certain that our individual acts are harmless or create no benefits in isolation in a given instant. This has a lot in common with Bernard Williams' (1981) argument about integrity: we should not be fragmented agents, but internally coherent. We might often fail, but overall, we should aim to live our lives in a way that corresponds with our values, and this includes what we are involved in as members or constituents of collectives. The idea is not to aim for some moral sainthood, or to fret over every dot in the picture, but to look at the patterns instead. Such patterns certainly form over months and years, but the relevant period of evaluation will vary. Although the literature on the problem of small contributions has focused on momentary choice situations, I am arguing that the central question should instead be individual behaviour over time.

By bringing up integrity, I wish to refer to the idea that we should not discount the effects that we bring about together when we think about how we should live our lives. Marion Hourdequin (2010) describes moral integrity as an obligation to avoid hypocrisy by accepting some level of personal obligation to try to fulfil a collective obligation one has accepted. After all, human psychology does not lend itself to stark separation between personal and political obligations, the individual and the collective. Although the exact concept of integrity is difficult to pin down, it includes both internalisation of certain commitments and unity among these commitments,

in other words, 'integrality' and 'integration' (Audi & Murphy 2006; Hourdequin 2010). I do not aim to offer reasons for caring about something for those people who in general do not care about it.[8] The point I want to make here is simple: that by discounting the collective outcomes we are contributing to, we could be letting go of some of our commitments as individuals. If one attempts to separate the individual and the collective domains starkly in moral matters, it can lead to a lack of coherence between one's values and contributions. Any blunt separation is an illusion. Such lack of coherence can also be harmful for the person's character.

While our everyday small choices affect the coherence of our moral lives, I do not aim to present an account that comes straightforwardly under virtue ethics. While the torturers could easily be covered by such an account – their willingness to play a part in torture pointing to cruelty or indifference to the suffering of others – it is harder to find such obvious character flaws in the more mundane small contributions to collectively-caused harms, especially when their roots are in structures. However, while character-based reasons are not the whole story, they are still an important part of the integrity account that I am proposing. If we willingly contribute to harm daily, however imperceptibly, we could become numb to the problem. Not caring enough can become part of your moral narrative if you start regarding such participation as morally fine, when in reality you are prepared to be part of a set of people who could blow up an island together.

The appeal is not to the aggregate impacts made by an agent over a period of time, but to the aggregate impact *on the agents themselves*: the corroding and corrupting effect on our characters if we fail to properly consider the impacts of the collective outcomes we contribute to. In time, such numbness to a problem can also contribute to creating harmful social norms around the issue, which can help to create the shared illusion that what we are doing together is at least acceptable, if not fine, even if the collective outcome is harmful.

With small contributions to collectively caused outcomes, it can be unclear what 'contributing' refers to, if it does not necessarily have a causal effect. The answer will depend on the kind of collective that is in question, that is, if the collective outcome is due to organised collective action or looser collective patterns of behaviour. To use *The Harmless Torturers* to illustrate the difference, the small contributions could be

[8] One might ask: what about those who don't have the relevant values? If I do not value the well-being of other people, or care about non-human nature, and I do not consider these when making decisions, then why should I care about the collective level features of my actions? The argument offered here does not seek to offer an overarching account of *why* we should care about small contributions. Its ambition is limited to trying to show that *if* we care about other people or nature more generally, *then* it is incoherent to not to also care about our small contributions to harms or good outcomes. This holds at the level of types of acts, but not with every token.

made in a setting of organised action (individuals who work for The Harmless Torturers Ltd.) or as a result of looser collective patterns of behaviour (e.g. the individuals are following some bizarre social norms).[9] In this article my focus is on organised collective action, whether by collective agents or more informal groups, but I have written about small contributions as part of looser collective patterns elsewhere (Hormio 2024: 91–97).

One example of an organised collective is the village council in *Rice Grains*. Christopher Kutz's (2000) notion of a *participatory intention* offers a simple and effective way of conceptualising collective action and helps in thinking through our individual responsibility as members of collective agents. We share a goal that teleologically explains our actions when our participatory intentions overlap (and we are sufficiently aware of this). That is, they are not explained in causal terms, but in terms of the purpose they serve. The villagers donate a cup of rice each to make the party risotto, because the village council has decided to help feed the neighbouring villagers with a weekly feast. In causal terms, one donated cup does not make or break the collective goal of serving the party risotto, but it can nevertheless be conceptualised as contributing to it, because the purpose the donation serves is to be a gift towards the risotto. Participatory intention is made up of a *collective end* (the object of a description that is constituted by the acts of many individuals), and the *individual role* (action an individual performs to promote a collective end) (Kutz 2000: 81). In *Rice Grains*, the collective end is the party risotto, and the individual role is to donate a cup of rice to it.

Individuals do not need to intend the collective end, as long as they interpret themselves as *contributing* to it. This account of joint action is more minimalistic than many other accounts, as it does not require a collective commitment, only an acceptance of collective norms.[10] We often act together under only a vague description of what we intentionally promote together (Kutz 2000: 155). Some of the villagers in *Rice Grains* might only contribute their cup of rice because their neighbours do, without any intention as such to feed people from neighbouring villages. Because participatory intentions need only sufficiently overlap ("contribute to the risotto"), the members do not have to intend every action that is performed for the collective end to count as members.[11]

[9] My interpretation is that the individuals are employed to be torturers or are otherwise members of an organised torturing collective. But if the dial was in a park with only an instruction note attached to it, with random people turning the dial, it would still be callous to do so. You would allow yourself to be involved in easily avoidable collective harm.

[10] Collective norms can be understood either as social norms particular to some collective, like an organisation or an association, or as behavioural regularities within the wider collective environment.

[11] Note that this does not rule out unintended collective consequences, but simply means that participatory intentions are directed at a goal that is intended. What we tolerate, desire, and value links us to the collective outcome. When it comes to accountability for collective action, the *basis* is individualistic (your participatory

One concern is that we need a reason to act (or refrain from acting) that connects appropriately to the collectively produced benefit or harm. A donation expresses support for the weekly feast and solidarity with fellow villagers, without having to make calculations about the expected utility of the contribution. On this broader conception of contributing, we can no longer explain why we should take one specific action rather than another (Nefsky 2015). If my action is not expected to make an actual difference to the outcome, but is meant mostly to express support and solidarity, then why not do something symbolic only? Since 99 cups of rice make enough risotto to feed all the partygoers, why should I contribute a cup instead of, say, singing about the virtues of donating rice? The problem, of course, is that if everyone thought like this when individual contributions are small, nothing significant would ever be achieved.

The answer to this lies in separating two worries: can an account tell us why we should take one specific action rather than another at a given instant, and can it explain why we should take that type of action in most cases? While there can be a reason to make or refrain from an action of a certain type, my account does not purport to be action-guiding in a choice situation, as it does not apply to all tokens of a type of action. I aim to give only a pro tanto reason; the all-things-considered reason depends on the circumstances. It might not even make sense to ask questions about contributions in all individual instances, and it probably does not matter what a given villager or islander does on a given day. My suggestion is that our evaluative focus should instead be on the patterns that our actions form.

Imagine a villager in *Rice Grains*, who often sings about the merits of contributing to the risotto, but week after week, fails to give anything. In my account, what matters is the narrative that forms over time. Or imagine an islander who displays posters outside their house about the importance of keeping an eye on water consumption but neglects to vote for the installation of water saving taps in nearby public amenities. In this case, an opportunity to make a difference is not taken. Symbolic ways can count as contributions sometimes, and can replace more concrete ways to contribute, but only as long as they are balanced by other acts as well. As the focus on the coherence of the moral life of a person is about the big picture, symbolic reasons of character matter alongside collective consequences.[12]

---

intention), but the *object* is collective (the outcome of collective action) (Kutz 2000: 115–116). Therefore, even if you do not intend to contribute to a collective harm, you can still be accountable for the bad outcome if you do not revise or question your participation.

[12] I should make it clear that my concern with character is not about trying to attain as clean hands as possible. Not only would that probably involve being somewhat of a hermit in the modern, interdependent world, but it would certainly result in lost opportunities to be part of a change for the better. My concern is rather that if we discount the results of the collective action that we contribute to, we are paying insufficient attention to a large part of our moral lives.

The idea of the coherence of the moral life of a person includes what we do and intend to do with others, and what we owe to them due to this. This does not mean that the only way to participate is through donating the rice. There might be weeks when singing or some other such expressive act makes sense. But if many others start also singing instead of donating their cups, the situation changes. When participating in collective action, one should always have their feelers out for changes in the collective context (Hormio 2024: 84).

Sometimes there are options to not contribute. Say that the islanders have dug new wells that are not connected to the Box of Doom. Unfortunately, the mad scientist dug her wells in places with the best groundwater reservoirs, so the new wells draw from more confined and unsaturated aquifers, with the result that the islanders must pump much harder to obtain water. They prefer using the old wells because they deliver more water with less work. Still, the new wells provide the islanders with an option to opt out of being part of the problem, albeit at a personal cost. This version of the *Box of Doom* resembles many purchasing choices in the real world.

The coherence of the moral life of a person does not usually stand or fall due to individual acts in certain time slices, as it is about the bigger picture and patterns.[13] In other words, an overly individualistic conception of one's footprint in the world is at odds with the interdependent reality of our moral lives as social beings engaged in multiple levels of (obvious and less-than-obvious) cooperation each day. Not applying the values we hold as individuals to our behaviour in collective contexts is incoherent, and in the long run, detrimental to the way we organise our lives together.

5. Conclusion

Sometimes the correct unit of moral analysis is not an individual act, but the coherence of the moral life of a person. Small contributions to a harm or to a good outcome matter if they are contributions to a normatively significant outcome, regardless of whether we can tease out any causal difference through a direct small choice or a direct small effect. Some contributions to collective outcomes operate under collective level structures that limit or pre-describe individual acts in such a way that there is no direct small choice or effect. To explain why individuals still have a pro tanto moral reason to make or withhold their contributions, the evaluation must encompass more than just the aggregate causal impact.

[13] If you have seen that the social norm around participation is robust, and that contributions are made week after week, it is not incoherent to contribute sometimes to the collective end in some way other than by donating your cup: you can be fairly certain that the collective goal is achieved regardless. Still, such alternative ways of contributing must always be made with the view of the collective context, i.e. the agent must monitor the collective context and stay alert to possible changes.

Although the focus has been on the impact of small contributions to an individual character, the goal is not to have scrupulously clean hands and perfect character. Rather, the goal is to become aware of oneself as enmeshed in several collective webs, an individual who is socially situated in an interdependent world. In other words, to drop an overly individualistic way of conceptualising the effects of one's actions. In collective harm cases, our evaluative focus should be on the patterns that our actions form with other such acts in a collective setting. Moral theorising should not try to isolate individual effects and agents in cases in which it does not make sense.

Our contributions, however small they might be, are part of our moral narrative. The collective goals we promote should be coherent with our values. Caring about small contributions forms part of a coherent moral life.[14]

References

Asker, A. S. (2023). The problem of collective impact: why helping doesn't do the trick. Philosophical Studies, 180, 2377–2397.

Audi, R. & Murphy, P. E. (2006). The Many Faces of Integrity. *Business Ethics Quarterly,* 16, 3–21.

Barnett, Z. (2018). No free lunch: The significance of tiny contributions. Analysis, 78(1), 3–13.

Broome, J. (2012). Climate Matters: Ethics in a Warming World. W. W. Norton.

Broome, J. (2019). Against Denialism. Monist, 102(1), 110–129.

Budolfson, M. B. (2019). The inefficacy objection to consequentialism and the problem with the expected consequences response. Philosophical Studies, 176, 1711–1724.

Collins, S. (2019). Collective Responsibility Gaps. Journal of *Business Ethics*, 154, 943–954.

Cripps, E. (2016). On Climate Matters: Offsetting, Population, and Justice. Midwest Studies In Philosophy, 40, 114–128.

Hormio, S. (2024). Taking Responsibility for Climate Change. Palgrave Macmillan.

Hourdequin, M. (2010). Climate, Collective Action and Individual Ethical Obligations. Environmental Values, 19, 443–464.

[14] I would like to thank Andrea Asker Svedberg and the workshop participants for very helpful comments.

Kagan, S. (2011). Do I Make a Difference?. Philosophy & Public Affairs, 39(2), 105–141.

Kingston, E. & Sinnott-Armstrong, W. (2018). What's Wrong with Joyguzzling?. Ethical Theory & Moral Practise, 21, 169–186.

Kutz, C. (2000). Complicity: Ethics and Law for a Collective Age. Cambridge University Press.

Lichtenberg, J. (2014). Distant Strangers: Ethics, Psychology, and Global Poverty. Cambridge University Press.

Nefsky, J. (2015). Fairness, Participation, and the Real Problem of Collective Harm. In M. C. Timmons (Ed.), Oxford Studies in Normative Ethics, Volume 5 (pp. 245–271). Oxford University Press.

Nefsky, J. (2017). How you can help, without making a difference. Philosophical Studies, 174(11), 2743–2767.

Nefsky, J. (2019). Collective harm and the inefficacy problem. Philosophy Compass. 14:e12587.

Nolt, J. (2011). How Harmful Are the Average American's Greenhouse Gas Emissions? Ethics, Policy & Environment, 14(1), 3–10.

Parfit, D. (1987). Reasons and Persons [reprint with corrections]. Clarendon Press.

Sandberg, J. (2011). 'My Emissions Make No Difference': Climate Change and the Argument from Inconsequentialism. Environmental Ethics, 33(3), 229–248.

Spiekermann, K. (2014) Small Impacts and Imperceptible Effects: Causing Harm with Others. Midwest Studies In Philosophy, 38, 75–90.

Wieland, J. W. & van Oeveren, R. (2020). Participation and Superfluity. Journal of Moral Philosophy, 17, 163–187.

Williams, B. (1981). Moral Luck: Philosophical Papers 1973–1980. Cambridge University Press.

Julia Nefsky & Sergio Tenenbaum [1]

Rescuing Ourselves from the Pond Analogy

Peter Singer famously argues that when we spend money on seemingly ordinary pleasures for ourselves, we are doing something gravely wrong. In the process, he (famously) draws an analogy between spending money in such ways and not saving a child drowning in a pond when you could easily do so. There have been many responses to Singer. Some of these make potentially important points and might give grounds for rejecting Singer's principles. But what they do not do, we argue, is respond effectively to the Pond Analogy and the argument it itself gives for Singer's conclusion. This reveals that Singer's focus on deriving his conclusion from general principles is a mistake; the hard-to-resist argument is the Pond Analogy itself. More broadly, we show that the Pond Analogy presents a crucial challenge to our ability to give a plausible, coherent conception of morality. We close by sketching our answer to it.

[1] Department of Philosophy, University of Toronto.

1. Introduction

Peter Singer famously argues that when we spend money on small pleasures or luxuries for ourselves, such as a dinner out at a restaurant or movie tickets or new clothes that we do not need to stay warm, we are doing something gravely wrong. Most of us regard these sorts of choices as ordinary, perfectly acceptable parts of life, at least insofar as they are not done excessively or extravagantly. But Singer argues that spending money in these ways is (typically) wrong. We ought instead to donate this money to organizations that provide life-saving aid to people in need. It is seriously wrong if we do not do so.[2]

The claim is not just that sometimes you should forgo some pleasures or benefits for yourself and donate the money instead. It is that (nearly) any time you spend money in these sorts of ways, you are acting wrongly, and seriously so.[3] Let's call this conclusion, "Always Donate" for short.

As is well-known, in arguing for this Singer draws an analogy between spending money on pleasures or luxuries for oneself and choosing not to save a child drowning in a pond when you could easily do so. He asks us to imagine the following scenario:

> **Pond:** On your way to work you pass a small pond, and you see that there is a young child drowning in it. There is no one else around. Wading in and rescuing the child would be easy and safe, but you would ruin your new shoes and suit.[4]

Of course, you ought to save the child in this scenario. If you don't save the child because you don't want to ruin your new shoes and clothes, this would be horribly wrong. But Singer's suggestion is that when you spend money on unnecessary pleasures or luxuries you are doing something equivalent to that. Instead of, say, spending $100 on a dinner at a restaurant, you could have an inexpensive meal at home and donate the remaining money to an aid organization. Doing so would - Singer says – save a life. Choosing to go to the restaurant, then, is just like choosing your new shoes over saving the child in the pond. It is choosing a small benefit for yourself over saving someone else's life. It is wrong for the same reason and to the same extent. This is the Pond Analogy.

There have been many attempts to reply to Singer, and some of these make important points. Our aim in this paper, however, is to show that various such seemingly promising responses do not actually reply effectively to the Pond Analogy, and to the

[2] (Singer 1972, 2019).

[3] Except, of course, if not donating is needed to preserve your mental health sufficiently, so that you will be able to make future donations, or if it would be in some other way counterproductive to donate on this occasion.

[4] Singer 1972, 231, and (Singer 2019, 3).

argument that the analogy itself gives for Always Donate. In particular, we argue that rather than explaining where the analogy goes wrong, these replies each require, for their own ideas to work as intended, *presupposing* that the analogy is mistaken. So, they do not actually dispel the analogy and the case it makes for Always Donate. Instead, they offer potentially important, good ideas which depend crucially, for their own functioning, on there being another independent way out of the analogy.

More generally, our paper aims to show that responding to the analogy is a *different* and more *fundamental* task than people have understood. We argue that we cannot get out of the Pond Analogy, and its implausible implications (like Always Donate), by developing a more plausible conception of morality than Singer's (and others like his), or by revealing that Singer's view neglects some important aspect of morality, or of how its demands interact with other aspects our lives. We show that it is quite the opposite: the development of a plausible conception of morality, and of the interaction between morality and our other ends and projects, *depends* on our being able to break the analogy in a prior, independent way.

How then can we get out of the analogy and its implications? In the final part of the paper, we sketch what we think is the answer. We propose that breaking the analogy requires recognizing the mistake in an auxiliary assumption in Singer's argument: his assumption that by donating money, you save a life. Many have just assumed that this assumption is true. But even when objections to it have been raised, people seem to think that our ability to get out of Always Donate should not hinge on that sort of point. We think, however, that this is, in fact, exactly where the mistake in Always Donate lies. Indeed, our view is that understanding why this sort of claim typically does not hold is essential to understanding the nature of our duties to help others, and the nature of imperfect duties more generally.

Before we begin, a clarification: as we've seen, there is a direct argument from the Pond Analogy to Always Donate. Namely, there is no morally significant difference between not helping in Pond and what we do when we spend money in the ways in question; therefore, since it is seriously wrong to not help in Pond, it is also seriously wrong to spend money in such ways. But Singer does not present his argument in that way. Instead, he is focused on giving an argument from general moral principles. The central general principle that he invokes is:

> **Singer's Principle:** "If it is in our power to prevent something bad from happening, without thereby sacrificing anything of comparable moral importance, we ought, morally, to do it." (Singer 1972, 231.)

This, he says, implies Always Donate when combined with a couple of auxiliary premises, which he takes to be uncontroversial:

(i) Suffering and death from lack of food, shelter and medical care are bad.

(ii) Donating to an aid agency, instead of spending money on a pleasure or luxury for yourself, prevents some such suffering or death.

Singer does defend the Pond Analogy, but he does so primarily in order to show that the principle applies, just as it does in Pond, to the decision of whether to donate money or spend it on a pleasure for yourself. That is, he defends it to show that there is nothing about that sort of choice situation – call it "Charity" – that brings us out of the ambit of the general principle. But while this is how he presents his argument, one lesson of our paper is that Singer's focus on the general moral principle is a mistake. The general principle that Singer invokes is easy to reject. The powerful argument is the direct one from the Pond Analogy itself. It is hard, we will see, to get out of that argument without begging the question.[5]

2. Agent-Centred Prerogatives

We are going to begin with Sam Scheffler's idea that morality includes an agent-centred prerogative. Scheffler doesn't develop this idea in response to Singer. But it provides a plausible way to reject Singer's Principle, and it might seem to explain why Always Donate is false. However, we will show that this does not work – not unless we can reject the Pond Analogy on prior grounds. We will then show that the same problem arises for ideas that have been given specifically in response to the Pond Analogy.

The idea of an agent-centred prerogative is the idea that there is a permission to give your own interests greater weight than those of other people. The contrast is impartial consequentialism, which says that we are required to always act in the way that will produce the outcome that is best from a fully impartial perspective. Scheffler points out that a very basic fact about human agency is that people do not operate from a purely impartial perspective. Each person has their own *personal perspective* from which they determine what they care about, evaluate how things are going, make decisions, and live their lives. "People do not," Scheffler writes, "typically view the world from the impersonal perspective, nor do their actions typically flow from the kinds of concerns that a being who actually did inhabit the impersonal standpoint might have." Arguably, morality must work with this very basic, core fact about human agency by to some extent allowing individuals to devote energy and attention to

[5] Singer points out that weaker versions of the principle work just as well for the argument. As we will see in section 5, this doesn't matter much for our point.

their own interests "out of proportion to the weight" they would receive from a fully impersonal standpoint.

Importantly an agent-centred prerogative is a permission to give only a certain amount of greater weight to your own interests. It is not a blanket permission to do whatever you want. So, it will not always be permissible to pursue your own projects or interests. But because there is some degree of permission to weigh your interests more heavily, this is supposed to make sense of how there is room for people to permissibly pursue, over time, their own interests, projects and relationships.

This might seem to give us a good way to reject Always Donate. If I can give extra weight to my own interests, this – we might think – explains why it can be permissible to not always donate money when I could do so at what, from an impartial perspective, looks like just a small cost. But does this work?

To be plausible, the extra weight the Agent-Centred Prerogative allows you to give to your own interests needs to *not* be enough to make it permissible to not rescue the child in Pond. The prerogative would allow you to weigh your interests in not ruining your new clothes more heavily than a fully impersonal calculus would. But this extra weight needs to not be enough to make it permissible to not save the drowning child, since you are certainly required to save the child at the cost of your clothes. Pond is exactly the sort of case in which the extra weight you can give yourself must not be enough to make it permissible to not help someone.

But this means that if the agent-centred prerogative is to be of any help in explaining why Always Donate is false, it needs to be able to apply *differently* in Charity than in Pond. Otherwise, we would have to say that in Charity, just as in Pond, while you can give your own interests extra weight, this extra weight is not going to be enough to make it permissible to go out to the restaurant, or to buy the new clothes, when you could instead donate this money and save a life. So, for the idea of an agent-centred prerogative to be of any use in showing the mistake in Always Donate, we need to be able to reject the Pond Analogy on prior grounds.

The point is not just that an agent-centred prerogative can't help us get out of Always Donate. It is also that it cannot do what Scheffler and others want it to do in general unless we can reject the Pond Analogy on prior grounds. Unless we can explain why helping in Pond is morally different from donating in Charity, an agent-centred prerogative will not actually be capable of justifying doing the things needed to pursue over time your major life projects, or to maintain your relationships, and so on. Suppose, for example, you have an aunt who you love and want to maintain a close relationship with. Can the prerogative explain why you are permitted to spend money and time on doing so? Well, it would be wrong not to save a child drowning in a shallow pond, even if you are on your way to your only chance to visit your aunt this year. So, unless we can say that the choice in one's actual life between visiting

your aunt and donating the time or money is *morally unlike* this variation on Pond, we would have to say the same thing there: that even with the bump from the agent-centred prerogative, your interest in visiting your aunt is not going to be enough to permit you to go ahead with the trip, rather than donate the money. For the agent-centred prerogative to be capable of doing what it is supposed to do – explain how it can be permissible to do things like visit your aunt, or take a philosophy class, or work on an art project (even when they are not optimal from an impartial perspective) – we need to be able to reject the Pond Analogy on prior grounds.

3. The Aggregationist Response

Several philosophers have proposed that the key difference between Pond and Charity is that Pond is anomalous, whereas the opportunity to donate money is constantly there. As long as you have some expendable income, you will always be in a position to help people living in extreme poverty by donating. This is a very different from what we assume to be the case in Pond – that it is a one-off, unusual situation.[6]

To get a morally analogous 'pond' case to Charity, we need to imagine that you are constantly encountering opportunities to rescue drowning children from ponds.

> **Constant Ponds:** In the city where you live there are ponds everywhere, and young children are constantly falling into them. It is impossible to do anything outside your home without coming across a child drowning in a pond. If you were to rescue every child that you could, you would never be able to get anywhere. Even when you are home, you know that an option available to you is to go outside and rescue children.[7]

Travis Timmerman argues that, while you must rescue the child in the original Pond, it is not true that in Constant Ponds you must always rescue a child whenever you could do so at only a small cost to yourself. He writes, "Few moral truths may seem more obvious than that one is obligated to sacrifice $200 to save a child's life at least once. But it's far from obvious that one is obligated, for his or her entire life, to constantly sacrifice everything comparably insignificant to a child's life." (Timmerman 2015, 211) According to Timmerman, for someone who spends much of their time saving children in Constant Ponds, it is intuitively permissible for them to, say, go to the theatre sometime, even if there is nothing major (like their sanity or their ability to provide for themselves) at stake in their doing so, and even though this means passing up an opportunity to rescue children.

[6] See (Timmerman 2015); (Thomson 2021); (Garrett Cullity 2004, 85; G. Cullity 2003); (Schmidtz 2000).

[7] There are a number of examples like this in the literature, but for this simple version see (Woollard 2015, 126).

We think it is actually very difficult to say what is permissible in Constant Ponds. But regardless, what's important for our purposes here is the suggested way out of the Pond Analogy. The suggestion is that whether a situation is anomalous or non-anomalous can affect what you may permissibly do on a given occasion. It is one thing to say you must rescue in a one-off, unusual encounter. It is quite another to say that you must do so each time in a series of repeated opportunities to rescue.

Why would it matter whether the situation is anomalous or not? Jordan Thomson proposes that it is because aggregate costs matter. When one faces repeated low-cost rescue opportunities over a long period of time, while the cost of any single rescue alone (looked at individually) may be low, the aggregate cost of performing all of the rescues is very high. Doing so would consume your life. These aggregate costs can justify refusing to rescue sometimes.[8]

Does this work as a way out of Singer's conclusion? To see the problem, return to the original "one-shot" Pond case. Regardless of the extent of your charitable contributions, you are morally required to rescue the child drowning in the pond in front of you. You cannot claim that since you do a lot to save lives at other times by donating to aid organizations, it is permissible to go ahead to the theatre and not stop to rescue the child. But how can the Aggregationist say this? Why don't the aggregate costs of your donations make it permissible for you to forgo helping this time? The Aggregationist wants to say that this is because Pond is anomalous. But *why* is it anomalous? For Pond to be anomalous it must be relevantly different from choices we face all the time. But this means that Pond is anomalous only if it is not morally just like the choice in Charity, since the choice in Charity is one that we face constantly. In other words, the very thought that Pond is anomalous presupposes that the Pond Analogy is mistaken: it presupposes that the choice in Pond is not morally just like the choice Charity.

Of course, Pond is, in certain descriptive features, an unusual encounter: most of us rarely encounter the opportunity to rescue someone from drowning specifically. However, the Aggregationist approach cannot merely appeal to these descriptive differences. Each opportunity to donate to a charitable organization is different in some descriptive features from every other. For instance, an opportunity to rescue someone by donating to the Ukraine war effort at this specific, current stage of the war is something you will never encounter again, and so is in that sense anomalous. For the Aggregationist approach to work, it must be that Pond is anomalous in some

[8] (Thomson 2021). See also (G. Cullity 2003). There might be a number of different ways of cashing out how this justification goes. Theron Pummer offers one attractive way of understanding it, using the notion of an agent-centred prerogative. Aggregate considerations, he argues, can amplify the prerogative that you have on a given occasion. See (Pummer 2023)

morally relevant respect – in a respect that distinguishes it morally from the opportunities we have all the time to help people at low cost.

So, the claimed difference between Pond and Charity depends on there being another prior answer as to why Pond and Charity are not morally the same. Thomson argues that the burden is on Singer to show that we must respond to each non-anomalous case just as though it were anomalous. But the very claim that Pond is anomalous assumes that the Pond Analogy is mistaken. The burden is really on us, therefore, to show that we can reject that analogy. Only then can we use the Aggregationist approach to explain the mistake in Always Donate. As with the agent-centred prerogative, the point is not only about responding to Singer. It is also that for this very plausible idea that aggregate costs matter to function as it is supposed to, we need to be able to reject the Pond Analogy on prior grounds.

4. Beneficence as a Duty to Adopt an End

A different approach that might seem to hold a lot of promise appeals to a broadly Kantian account of the duty of beneficence as a duty to adopt an end. Several philosophers have argued that this view can explain why in some cases – like Pond – one must help in the particular instance, while in others – like our ability to donate to charitable organizations – there is room to choose when and how to help. Let's start with (Stohr 2011)'s version of this view.

On the Kantian view, the duty of beneficence is a duty to make the happiness of others one's end, and Stohr argues that this generates two related duties: a wide duty to help others, and a narrow duty "to avoid indifference to others as setters of ends" (Stohr 2011, 61). On the wide duty: you do not need to help everyone at every opportunity to count as having the happiness of others as your end. But you cannot count as having this end if you never help anyone. So, you must do some helping, but there is room to choose when and how. However, in addition to this wide duty, having the happiness of others as your end requires never being indifferent to others as setters of ends (the narrow duty). I manifest indifference to someone, roughly, when I treat their ends as not worth taking into account in my deliberations. If you ask me for the time and I just keep walking without even acknowledging your request, I have manifested indifference to you. On Stohr's view, "although we are not always required to help, we are always required not to be indifferent." (Stohr 2011, p. 45)

Often I can avoid manifesting indifference without helping you pursue your end. I can show that I care about your plans to go on vacation by just expressing joy about it. I do not need to help you realize your plans by, say, making a financial contribution to your travel costs. Or if you are moving, I do not necessarily need to help with the move in order to not manifest indifference. Telling you that I am busy that day and

thus regrettably won't be able to help is often enough. However – Stohr says – in some situations helping someone is the only way to not be indifferent to her. In such a case, one is required to help. In Pond, politely explaining to the drowning child that, regrettably, I have dinner reservations and won't be able to pull her out of the pond does not suffice to count as having an attitude of recognizing the child as a setter of ends; I must rescue the child.

On Stohr's view, the key difference between Pond and Charity is that there is no way to not manifest indifference to the drowning child *other than* to save him from the pond. This is why it is obligatory to do so. Whereas, there are other ways besides, say, forgoing this particular trip to the movies to avoid manifesting indifference to the global poor. You can do so by donating or volunteering at other times. When helping someone is the only way to not be indifferent to them - as it is in Pond - you are required to help. When it is not the only way, you have latitude to choose when and how to help - as with Charity.

But what allows us to say that Pond is on one side of this divide and Charity is on the other? *Why* does it not express indifference to the victims of poverty, famine, war etc., for me to go out to the movies or a nice dinner when, if I just donated this money instead, I could save one of their lives? To say that it does not express indifference seems to require that we already see that it is a mistake to regard the choice between spending money in such a way and donating it as just like the choice in Pond. Insofar as it looks potentially *just like* the choice in Pond – a choice between a pleasurable evening and saving someone's life – it will seem like not donating in this particular occasion does manifest indifference.

Now Stohr does say more to explicate the difference between instances where helping is obligatory and those in which it isn't. Let's consider if this elaboration solves the issue. She writes:

> The cases where refusing to help is most likely to be obligatory seem to be those in which it is reasonable for the other to expect *me* to help, and where there is considerable discrepancy between the need I could meet and the costs I would incur by helping. In such cases, I disregard an expectation of help that is reasonably directed at me by another rational agent without having anything plausible to offer as an excuse for not helping. And that is what expresses the prohibited indifference toward others as setters of ends. (Stohr, 2011, p. 64)

When we look at the choice between going out to the movies and donating the money, the condition of considerable discrepancy is clearly there (at least as long as we assume, with Singer, that by donating, you would save a life, and Stohr does not question that assumption.) So whether this account works to explain why going to the movie does not express indifference comes down to whether we can say that 'the

other' could not reasonably expect me to help them. It might seem easy to say that. How could victims of poverty in general, or the particular victim(s) that my donation would help, reasonably expect *me* to help? They do not even know I exist.

But this cannot be the relevant difference. Imagine the child in Pond is unconscious, or otherwise unable to see that I am there. They, then, could not reasonably expect that I help them. But this does not release me from the obligation to help. One might reply that, while subjectively the child wouldn't have that expectation, there is an objective sense in which it is reasonable for her to expect that I help. Were she to know that I was there, the expectation would be reasonable. But why is not the same true for the potential beneficiaries of my charitable contribution? Why is it not, in this objective sense, reasonable for these people to expect that I would help them?

Stohr might point out that, unlike in Pond, I am not uniquely positioned to help them. But this again, can't be the difference we need. Consider the following case:

> **Crowded Beach:** There are fifty people on a beach. A boat capsizes close to shore, and six children start drowning. Each of the fifty are capable of easily rescuing at least one of these children. So, if at least six of them go to help, all would be rescued. You are one of the fifty.[9]

Here, while you are not uniquely positioned to help, if not enough others are acting so as to rescue all the children, you are certainly obligated to rescue at least one. Even if you will only be able to save one child, and thus no individual child can expect you to save them in particular, you still must go in and save someone. So, neither being uniquely positioned to help nor there being a particular victim who can expect you to help them specifically can be required.

In sum, without a prior account of why Pond and Charity are relevantly different, there is no understanding of "reasonable to expect" that applies to variations on Pond in which there is a duty to help, but that does not apply to my ability to donate instead of spending on a relative luxury for myself. Thus, again, the claim that in Pond, but not in Charity, I manifest indifference by not helping *requires* a prior, independent explanation of the relevant difference between the two scenarios.

A similar point goes for other Kantian accounts. Noggle, for example, takes the duty of beneficence to be the duty to have the well-being of others as one of our "ultimate ends" (Noggle 2009, 8). Ultimate ends are our most fundamental ends: ends whose pursuit is a "fundamental part of [a person's] life and identity" (Noggle 2009, 8). These ends are salient in an agent's deliberative field; they play a significant role in her choices and decisions. On this view, a person fulfills their duty of beneficence by having beneficence as one of their ultimate ends. The duty does not require that

[9] (Igneski 2006). See also (Singer 1972, p. 233).

beneficence is our only ultimate end, nor does it require that we always sacrifice other ends, or even trivial pursuits, for the sake of it. What matters is that the pursuit of beneficence is "a *central* project of one's life" (Noggle 2009, 11). This explains why there is room to choose when and how much to help others, including with our donations. But why, then, is it wrong to not rescue the child in Pond? Why can't I choose to pursue a different one of my ends at that moment, as long as I help others enough at other times? Noggle's answer is that Pond is a 'golden opportunity' to pursue the obligatory end; it is a special sort of situation in which not taking the opportunity would indicate that you did not really have beneficence as your ultimate end. But, just as with Stohr, the idea that Pond, but not Charity, is a golden opportunity requires a prior understanding of why charitable giving is not just like rescuing in Pond. Insofar as the choice in Pond and the choice in Charity seem morally the same – a choice between saving someone's life and a small benefit for yourself – there will not seem to be any grounds for counting Pond on the 'golden' side of this divide and Charity on the other.[10]

5. The Pond Analogy and the Project of Understanding Morality

Something that the views we have discussed all have in common is that they each reject Singer's Principle in some way or other. The ideas they advance are, if they work, reasons to reject his principle. But one thing we are seeing is that rejecting Singer's Principle does not necessarily do anything to get us out of Always Donate. It will not help *unless* we can reject the Pond Analogy, and these views do not tell us how to do that (in a non-question-begging way).

You might think that, of course, rejecting Singer's Principle does not necessarily help; as Singer's himself points out, weaker versions of the principle suffice for Always Donate. (e.g. 'If it is in your power to prevent something *very bad* from happening, without sacrificing anything *nearly* as important, it is wrong not to do so.'[11]) But our point applies to weaker versions of the principle too. Just as with the stronger version, arguing against the weaker principle won't help get us out of the pressure toward Always Donate, *unless* we can reject the Pond Analogy in a prior, non-question-begging way. This is because even if we take issue with the general weaker principle, we still must recognize that it would be wrong not to rescue the child in Pond. So – whether or not the general principle is true – insofar as the choice in Charity looks

[10] See Igneski (2006) for another interesting Kantian proposal. The same sort of argument we make in the case of Stohr and Noggle also applies to Igneski's proposal.

[11] Combines the weaker formulations from Singer (1972) and Singer (2009).

just like the choice in Pond (roughly, a choice as to whether to save a life at very little cost to oneself), there will remain pressure to say that what is true in Pond (namely, you must help) also goes for Charity.

More broadly, the lesson of our discussion thus far is that rejecting the Pond Analogy is a different task, and one that is more *fundamental* to the project of understanding morality, than people have realized. A plausible conception of morality must make sense of how the demands of beneficence can co-exist with some degree of space to pursue our own lives, and the relationships, projects and pursuits we care about. Singer, and others, have threatened the idea that there really is that sort of space, given the suffering and need that exists in the world. But while we might take the tremendous suffering and need to give grounds for thinking that our duties of aid are much more demanding than people ordinarily suppose, this is not the same as accepting that there is little-to-no room for pursuing any other ends. It is also not the same as accepting that we act *gravely wrongly* when we do things like go to a concert, or buy a new shirt, or eat at a restaurant. Those conclusions remain highly implausible. Thus, many people have attempted to develop and motivate a different sort of conception from Singer's (and others like his) of how the demands of beneficence work. They've tried to develop ideas which reject principles like Singer's, and which are intended to explain how the demands of beneficence (even if demanding) can co-exist to some extent with other projects and aspects of our lives. The Agent-Centred Prerogative is one such idea. Another is the Kantian idea that beneficence is a duty to adopt the happiness of others as *one of* your ultimate ends alongside others. The idea that aggregate costs matter is another contribution to this general project. And there are more. But what we have been seeing is that the Pond Analogy threatens the coherent intended functioning of these sorts of ideas. These ideas cannot really be said to do what they are supposed to do without a prior, independent rejection of the Pond Analogy.

Reflection on cases like Pond shows that the demands of beneficence can disrupt the pursuits of any of our ends, from the most frivolous activity to the most central project in our lives. Faced with the opportunity to save someone's life in such a situation, only quite major sacrifices or risks (e.g. one's limbs, one's own safety, perhaps one's long-term well-being) seem capable of justifying not helping. I have to rescue the child even if I am on my way to my only chance to visit my beloved Aunt this year. Or, I have to rescue the child even if my dream is to become a successful actor, and I'm on my way to an important audition. Any view needs to accept something like these conclusions if it is to be plausible. So, for an idea or apparatus to make sense of how there is a fair amount of moral space to pursue one's own life, it must be able to treat the opportunities we have *all the time* to help others in need (by donating money, by volunteering, etc.) very differently from the sort of opportunity we have in Pond. Without grounds to treat them differently, it will not be capable of

actually explaining how we are permitted to do things like visit our Aunt, or take bassoon lessons, and so on. It will not, in other words, be able to explain how we are permitted to do the steps or activities along the way that constitute our pursuing our projects or maintaining our relationships, because it will not have the grounds to treat the choice about any such step differently than it treats the choice in Pond. So, these ideas or apparatuses depend, for their ability to do what they are supposed to do, on a rejection of the Pond Analogy. And since, as we have been seeing, they do not themselves tell us where the analogy is mistaken, this means that they depend on our being able to reject it on prior, independent grounds.

Thus, conceptions and ideas of how the demands of beneficence are consistent with some space to live our lives *require* a prior resolution of the Pond Analogy. These ideas do not themselves tell us where the analogy goes wrong. On the contrary, for such ideas to be capable of counting as explanations of what they are trying to explain, and for them to be able to function as they are intended to function, they require a prior dismantling of the analogy.[12]

6. Answering the Pond Analogy

What then is the way forward? Singer's Principle and its variants have received great scrutiny in the literature that followed his seminal paper; yet his auxiliary hypotheses have largely been left unchallenged. Philosophers have sometimes pointed out that the assumption that each donation saves a life is naïve or empirically dubious,[13] but most have nonetheless effectively granted the assumption – either because they assume that, even if its oversimplified, it sufficiently approximates the truth, or because they think being able to show that the Pond Analogy and Always Donate are false should not depend on this sort of point. They want to be able to show these strong conclusions are false *even if* the assumption is true. In this section we argue that this is a fundamental mistake. The Pond Analogy fails exactly because that assumption is false. Your difference-making potential in Charity, contrary to Singer, is nothing like your difference-making potential in Pond or Crowded Beach. Getting clear on this is the correct and crucial way to break the analogy. Indeed, we think this is crucial to understanding the very nature of the duty to aid in general, and – even more generally – to understanding the nature of all imperfect duties.

[12] The positive flip side of this, though, is that if we can reject the Pond Analogy on prior grounds, various other supposed issues with such conceptions clear up. Various supposed problems for these views go away once we break the pond analogy properly.

[13] See (Temkin 2022) for a recent example.

6.1 Difference-Making and Collective Efforts

Let us suppose that Doctors Without Borders (DWB) makes an appeal to help the victims of an earthquake in Turkey and I am considering whether I should contribute, say, $100, to it. What exactly will my contribution do? DWB is not waiting for my contribution to start operations. It is also not as though a DWB volunteer is waiting on the phone with a pharmaceutical company, and they keep updating the orders as contributions come in. Instead, they will probably plan operations that are compatible with their expected budget and will deal with potential budget shortfalls as they come along. Realistically, my individual $100 contribution will not make a substantial difference to what DWB does in Turkey. How DWB proceeds on the ground will not go substantially differently give or take my contributing $100.[14]

This is not to say that my contribution of $100 is useless, that it doesn't help DWB's efforts. When it comes to making a contribution to a collective effort like this, whether, or how much, the effort will go differently depending on your individual contribution is, we think, not the only factor in determining whether your contribution is helpful, or how helpful it is (Nefsky 2017). So, this is not to say that it is not truly worthwhile and good to donate the money. But it does tell us that giving this money to DWB is unlike Pond in a crucial respect: it is unlikely to make a substantial difference itself. And, in particular, one such donation is very unlikely to make a difference on the scale of life or death to someone.

Something similar is true even if I contribute not financially but by flying to Turkey to be one of the doctors in DWB. Of course, I'll be treating particular people once I'm there, and once I arrive I'll have a duty to do the tasks I am assigned to. This work is – in our view – very important and helpful. But still, when I am deciding whether or not to go, I may have no reason to believe that things would go substantially worse were I not to do so. I may be confident that someone else would take my place or pick up the slack if I didn't do it. This, again, is not to say that my volunteering in this way is not extremely useful and helpful, or that there is not very good reason to do it. The relief efforts would be seriously compromised if no one would be willing to volunteer their time and skill, and again, once I am there I will be doing important work. The point is just that there are often no grounds for thinking that things would go substantially worse for those in need of aid were I to not take on the role.

In quite stark contrast, in Pond, I clearly will make a substantial difference. There is a particular person who my actions will save. There is someone who will live who would not otherwise have. Even if I am not sure that I will be able to save them, I at least know that I have a substantial chance of doing so. It is not only that I know that

[14] See (Garrett Cullity 2004; Temkin 2022) for similar points.

I will or will likely make a difference. It is also that it is clear what sort of difference I could make with my intervention: the difference between life and death for this person. This is an important, morally significant difference between Pond and Charity: in Pond, and also in Crowded Beach as described above (in which not enough others are helping to save all the lives), my helping action has clear, strong individual difference-making potential. The same is not true of my donations to charitable organizations, or even my volunteer work. While such contributions help the causes they are aiming to contribute to, it is not the case that we can say that things would, or are likely, to go worse in clearly definable, substantial ways were we to not make any given such contribution.

Here is a variation on Pond that is more analogous to Charity in the respect we are identifying:

> **Collective Effort at Crowded Beach:** Due to an unexpected tide change, several children are at risk of drowning in the ocean by a crowded beach. A group rescue effort is already underway, with many capable, equipped people already involved. For whatever reason (e.g. a couple of the children are further out in the water and time is running out), it is doubtful that the collective effort will be enough to save all of the children. But what is, justifiably, apparent to you is that, given your own limited capabilities, and given the collective effort already underway, your involvement will not make any substantial difference to what happens. You do not think getting involved would negatively impact the group's efforts, but what you can tell is that the outcome, and the effort involved in getting there, will not be substantially different give or take your joining in the rescue. And, in particular, adding yourself into the mix is not going to make the difference between life and death for any of the children.

This example is close to analogous to Charity along the dimension we are specifying: your individual difference-making potential. Is it wrong to not go into the water and try to help in this scenario? While many of us might be strongly inclined to try to do something helpful in these circumstances, and while that is a very good thing, it is plausible that someone who looks on to the rescue effort from the shore with great concern, but does not join in the effort has not acted wrongly. If it is clear to them that their involvement would not make any substantial difference to the success of the effort, it seems permissible for them to trust the rescue effort to these others who are engaged in it, and to carry on with their plans (e.g. to meet a friend, or teach a class).

The key difference between Pond and Charity is the same as the difference between Pond and this example: in Pond you will or could make a clear, large difference (a difference between life and death for the child) while in Collective Effort at

Crowded Beach and in Charity you will not. This explains the difference in our duties. It explains why it is not wrong to pass up a particular opportunity to give to a charitable organization.

Importantly, this is not only true, but it is already operative in ordinary thought. When people pass up a given charitable-giving opportunity, they do not typically think of themselves as choosing to let someone die whom they could save but with permission to do so. When, say, a canvasser comes to my neighbour's door collecting money for a relief organization, and she turns them down, she is probably not thinking "sure, someone is going to die who I could have just saved, but that's fine because they are far away", or "but that's fine, because I save other people at other times, and I can't be expected to save everyone". She is, most likely, not thinking of the stakes of her choice as life or death for someone at all, or anything else similarly large. My neighbour seems to operate under some awareness (even if only implicit, unarticulated) of exactly this fact: that, while the charitable organization may be doing very important work, and while contributing to it may be a very good thing to do, no one's life (or something similarly important) hangs on one such particular decision of hers.[15]

Of course, people do often say that the reason to donate is that doing so will *save lives* or will *make a difference*. But people often say such things without meaning, or having any clear impression, that a single donation will save lives that would not otherwise be saved, or that it will make some other sort of substantial difference in itself. They typically mean something much looser than that. For instance, they simply mean that the charitable organization is engaging in life-saving efforts and that one should donate to pitch into these efforts. This is most clear when one contributes to a very specific fundraising effort – say, raising money for an urgent operation for an uninsured asylum seeker. As I add my contribution, I might have no doubt that the fundraiser will hit its target; in fact, I might see that there are still days to go in the drive, and post the last $100 needed knowing full well that if I didn't post it someone else would. My motivation here is clearly to help in the collective effort, but it is clearly *not* to make a large difference. I know that things will not go very differently for the asylum seeker give or take my making this contribution.[16] Various charitable-giving opportunities are different from this in that there may be no final target, and no expectation that the organization will receive enough to do all that it could do.

[15] Certain things – like reading Singer, or perhaps watching certain ad campaigns – can muddle us on this. We then may start grasping for other explanations as to why it's fine to pass up a given chance to donate. But most of us return, in our day-to-day (non-academic) dealings, to thinking about such decisions quite similarly to how we did before.

[16] It is possible that my contribution will make a small difference – say, ease the asylum seeker's stress a bit sooner. But this is not my motivation for donating. And even if it is, this is not the sort of difference-making potential that Singer is talking about or that would obligate me to donate.

But still, the motivation to donate is, typically, quite similar: to help in the collective effort – to contribute to advancing its cause – without any impression that something large (like someone's life) *depends* on your particular contribution. Our point is that we are right to think that way.

Even if you agree that there is typically the difference we describe between Pond and Charity, you might be concerned that this is just a contingent feature of charities. Couldn't charities be organized such that my donation clearly does make the difference between life and death for someone? If so, wouldn't our view be unable to explain why there is no obligation to give at each opportunity in which one could do so at little or moderate cost to oneself?

Imagine how this might go. Imagine there is a Charity that sets things up so that people in need really are hostage to your particular donation.[17] Imagine "No Scruples Against Poverty" (NSAP), which sends a picture of a different baby to each potential donor with the following threat: "if we do not receive your contribution by Friday this baby won't be able to receive essential medical help and will perish". They assure you that they are not just trying to make vivid the importance of the work your money would be helping with; this is really how NSAP operates. Once they send you a request with a picture of a baby, they'll only use *your* money to help the baby in the picture, and if your money does not arrive, the child will be left to die. Igneski points out that such a tactic would be deeply immoral, and we agree with this. NSAP is manipulating donors in a way that wrongs the babies by holding their lives hostage to particular donors' willingness to donate. But whatever you think about NSAP's tactics, your moral situation has changed.[18] You can no longer relate to NSAP the same way you relate to other charities, and it seems now that they did succeed in putting you under a serious pro tanto obligation to send the money. This is indeed a consequence of our view, but it seems to us a plausible consequence.[19]

6.2 Unlikely Rescues

One might think, however, that it is not true that things would need to be very different than they currently are for your donation to have at least the *potential* to make a substantial difference (on the scale of life or death for someone). After all, isn't there some small chance that by donating to an effective charitable organization, your donation could make the difference between life and death for someone? Of course, Pond is not a scenario in which there is *just* a small chance of saving someone. However,

[17] Adapted from an example considered by Igneski (2006).

[18] Igneski agrees with this point as well.

[19] The fact that NSAP has an immoral charity design, might make you pause before accepting that you must contribute. But it does not change the fact that the structure of your obligations have changed.

one might argue that one need not accept the original analogy to have a powerful argument for a very strict duty to donate in Charity. Instead of Pond, one can use an analogy to the following example:

> **Unlikely Pond:** as you're going to work you see a child drowning in a pond. You realize that they've been in the water long enough that there is only a faint hope that they are still alive. By wading into the pond and pulling them out, you are highly unlikely to save them, and doing so would ruin your new expensive shoes.

Intuitively, even in this scenario, you would have an obligation to sacrifice your nice shoes to try to save them. This reply in fact grants our main point so far: Pond and Charity are not analogous because one's difference-making potential is very different in those two scenarios. However, it puts forward a similar argument for Always Donate: since Charity *is* analogous to Unlikely Pond, and you do have a duty to try to save the child in Unlikely Pond, you similarly have a strict obligation to donate whenever you could do so at small cost to yourself. Of course, this is a less powerful argument for Always Donate: it is much less implausible to deny that there is a perfect duty of rescue in Unlikely Pond than it is to deny it in Pond. But, still, we agree that we would have such a duty in this new scenario. However, in the remainder of the section, we will argue that this revised analogy also fails; Unlikely Pond is also not analogous to Charity.

There is indeed a small chance that your donation could make the difference between life and death for someone. Maybe your donation will, for example, make the difference between a shipment of supplies going out a bit earlier or later, and maybe this will make the difference between life and death for someone. This is extremely unlikely, but it's possible. But this sort of remote possibility does not make Charity similar to Unlikely Pond. In the typical case of donating to a large aid organization, there aren't grounds for thinking that this sort of remote chance is any greater than the remote chances of the reverse happening: of your donation actually making it the case that one fewer people are saved. It could be that in virtue of your $50, a fundraising agent for the charitable organization does not feel the need to do more fundraising that day (e.g. because they've already reached the quota they were aiming for), and thus they miss recruiting what would have been a substantially larger donation. This could have the downstream effect than fewer people are helped. Or, it could be that if you hadn't given the $50, some higher executives in the organization would have, upon examining the numbers at their next meeting, deemed donation revenue to be a bit too low, and decided that there is the need for a new fundraising campaign; it is possible that had that happened, many more lives would have been saved. These possibilities are very remote. But there is typically no reason to think that possibilities

like these are any more remote than the remote possibility that your $50 donation would change the order or scale of the organization's operations in such a way that they save more lives than they otherwise would have. What is likely is that nothing large (nothing like whether someone lives or dies) is going to turn on your individual donation. And we typically have no grounds for thinking the remote chances that something large does turn on it are more positive than negative. Because of this, these sorts of remote chances can typically be reasonably ignored. The choice in Charity should not be thought of as a choice as to whether to take a small chance of saving a life (as in Unlikely Pond). It should, rather, be understood as a choice as to whether to help in a collective effort aimed at saving lives or preventing suffering, where – while your contribution would be helpful – nothing big is going to turn on it.[20]

However, let us suppose, for the sake of argument, that there is specific reason to believe that there is a small chance that someone's life will be saved in virtue of your donation, and that there are no comparably-sized small chances running the other way. Even on this assumption, the choice in Charity is significantly different from the choice in Unlikely Pond. This is because there is a different disanalogy between Pond cases and Charity that is relevant as well: in Pond and in Unlikely Pond there is a particular person whom my reason to help picks out. The content of my reason, or obligation, to act is that doing so will save *this particular child*. When it comes to donations, on the other hand, there are no particular people who figure – as particular individuals – into the reason to give. Of course, in donating, I am trying to help, or contribute to helping, someone or some people, and if people are helped, these people will be particular people. But still, these individuals do not figure as particular individuals into the content of the reason to help. The reason in this case is something like "so that I will help, or contribute to helping, someone or some people". The reason to try to help the child in Pond, on the other hand, is something like "so that I will help *this child*".

This difference is significant. It is an important feature of non-consequentialist views that different persons, unlike different life-stages of persons, cannot be simply traded in our moral consideration. On these views, reasons or duties that concern particular people are different in content from reasons or duties that concern other particular people or that do not concern anyone in particular at all. If I am aware that my friend Larry is in serious need of help and that I can help him, I have a duty to help *Larry* in particular. This duty to help Larry is different in content from my duty to help my friend Mary. And these two duties cannot be treated as simply indifferent instantiations of a duty to help a friend. I cannot think that I equally comply with my

[20] We develop this argument in more detail in Nefsky and Tenenbaum, "Expected Utility Arguments and Tiny Chances" (in progress).

duty to help Larry if I help Mary instead, or that I comply better with the duty to Larry if I help two other friends instead. This is not to say anything about the stringency of these duties, or how different pro tanto duties compete with one another. The point is that an obligation to help *someone* is different in content and nature from an obligation that concerns a *particular person*.[21]

So situations in which there is a particular person who I must decide whether or not to try to help are relevantly different from situations in which I must decide whether or not to try to help *someone*. This difference may not ground a permission not to rescue someone when we *know* that a very small sacrifice will suffice to save their lives. Imagine a scenario in which the difference-making potential of giving up one's expensive shoes is as Singer mistakenly supposes it to be in Charity. Perhaps a virus has escaped from a lab, and it is certain to infect exactly one person, but there is no fact of the matter yet as to who this person is. I get a text explaining to me that, whoever this person ends up being, their only chance of survival is my immediately flinging my shoes into the lake. Once my handcrafted shoes touch the unique glacial lake in front of me, a rare chemical substance will be released that will quickly find the virus and kill it.

In this scenario, I am under an obligation to fling my shoes into the lake. Here I clearly will save a life. Even if I have no idea whose life it will be, and so there is no particular person who figures into my reason to give up my fancy shoes, the fact that I will save *someone's* life is enough to obligate me to take this opportunity to help.

But the revised analogy only claims that there is a *small chance* that your donation could make the difference between life and death for someone. Plausibly, I do not have an obligation to do things that have only a small chance at saving some life whenever I could do so at small cost to myself. For example, suppose I could, at a relatively low cost, buy a small defibrillator to carry around with me; I am not obligated to do that even though it affords me a small chance to save someone's life. A situation, though, in which there is a *particular person* whom you must decide whether to take a small chance to save is morally distinct from this. The reason at play there is different in content and nature, and could obligate you even if a similarly small chance of saving someone or other would not.

In fact, this difference can be deployed by the accounts given in sections 2 – 4 in order to explain why there is a relevant difference between the small chance of saving a life Unlikely Pond and the (supposed) small chance of saving a life in Charity. Let's look at the Aggregationist approach as an example. When I spend some money buying, say, some new fashionable clothes, I could have instead – we are supposing – spent the money so as to take a small chance at saving someone's life (by donating it).

[21] Tenenbaum argues for this in detail in (Tenenbaum 2024).

But that option is constantly there, since I always have the option to donate. I constantly have the option to act on the reason *there is a small chance of saving someone's life*. If I acted on this reason on every occasion when the costs of doing so, taken in isolation, would be low, the overall costs would aggregate so as not to allow me to pursue any of my other ends or projects. So, these aggregate costs can explain why there isn't an obligation to always act on this reason; there isn't an obligation to always act on the reason 'there is a remote chance of saving someone's life'. In Unlikely Pond, however, the content of your reason to help is different: it is a reason to try to save *this particular child*. The reason to try to save this particular child is not the same reason in content as the reason try to save someone. And I don't expect to go through life with an overwhelming number of occasions of having a tiny chance of saving this particular child. This is why the situation in Unlikely Pond is anomalous, and we cannot appeal to aggregative costs in considering the extent of the obligation to respond to this reason.

7. The Duty of Beneficence and Imperfect Duties

None of this is to say that there isn't an obligation to give to charitable organizations - the duty might even be quite demanding. Our aim was to identify a morally significant disanalogy between Pond and Charity that will allow us to explain why Always Donate is mistaken: why it is not wrong to pass up an individual opportunity to give to a charitable organization for the sake of a small pleasure or luxury for oneself. Our view is that the key morally significant difference is that your choice in Charity – unlike in Pond or Crowded Beach – is *not* one of potentially making the difference between life and death for someone (or anything similarly large). It is, instead, a choice as to whether to help a collective effort aimed at saving lives or preventing suffering, where – while your contribution would be helpful - nothing big is going to turn on whether or not you do so. When a contribution can help, but nothing major (nothing on the scale of someone living or dying) is going to turn on it, it is not wrong to refrain on any given occasion. But given that many people making such contributions plays an important role in funding the life-saving projects of aid organizations, we do think there is an imperfect duty to make such contributions: a duty that one satisfies by contributing enough over time.

It is not just our duties to give to organized charities that have the structure just described. Suppose I stop on the way to work in order to help someone whose car battery needs a boost, or to help someone who seems to be having a bit of a hard time carrying a heavy load into their building. Or suppose I hand a tissue to someone who has been crying on the subway. The reason to do these things is *not* that something major – e.g. some substantial suffering, or someone's life – might depend on my doing

so. It is not, in other words, that something important *turns* on whether or not I do it. Often this is because there is simply nothing large at stake. Other times it is, or is also, because I can be confident that if I did not come to the person's aid, someone else would do so. Either way, because of this, any one such act is typically not morally required. It would not be wrong (at least not in itself) for me to not dig into my bag for the tissue to hand to the stranger, or to not help the person struggling a little with their heavy parcel. (If you had reason to believe that something major could turn on your doing so – for instance, if the person looks like they might seriously injure themselves without help – this would be different.) Still, while any one such act is not morally required, we think it would be wrong to never do such things. We can think of ourselves as having a duty to take part in a collective effort to help each other. There being a general supply of good will – people inclined to help and be kind to each other, even when nothing much is at stake – is part of what makes life good. Without a general supply of good will our lives would be greatly impoverished. Because of this, we can think of human beings as having a collective duty to ensure that there is a sufficient supply of good will. While I am not obligated to do any particular helpful act when nothing major turns on it, doing some things like this sometimes is not merely optional and supererogatory. It is wrong to never do such things. We have an imperfect duty to help in these sorts of ways. This is – we think – how to understand beneficence as an imperfect duty. And as we have indicated, not all of beneficence is like this: when we are in a situation in which something major does, or does likely, turn on whether or not you help (as in Pond) the duty is then a perfect one.

Indeed, in our view, something similar is true for all imperfect duties. Not all imperfect duties connect to collective efforts or collective duties, and some of our imperfect duties are directed duties – duties to particular individuals. But the source of the duty's imperfection is, we think, the same in all cases. Take our duties to our children. There is a perfect duty to provide for their basic survival needs. But there is also an imperfect duty to foster their happiness and contribute to their development into thriving adults. The imperfection of this duty is, we think, grounded in the same sort of facts as it is with beneficence. Typically, any one thing I might do to contribute to their happiness or development is not in itself crucial: it would, typically, be a mistake to think that anything major *turned* on any one such act, no matter how nice, helpful or good it might be. But doing enough such things, regularly over time, certainly does make a big difference to their overall happiness and development, and this is what we are obligated to do.[22]

[22] For a different example: take the duty of gratitude. The duty of gratitude shares some very general feature with the general duty of beneficence, at least in case of a significant benefactor: no particular action expresses gratitude on its own, and the contribution of each particular action (except for "golden opportunities") is at best incremental. Of course expressing gratitude for a very specific thing (you pick up my wallet that fell on the ground)

Acknowledgements

This work was supported by SSHRC Insight Grants. We are very grateful for comments and discussion to Gunnar Bjornsson, Jordan Thomson, Tom Hurka, Rutger van Oeveren, Theron Pummer, and audiences at McMaster University; Colgate University; the University of Toronto Centre for Ethics; Upsala University; the Collective Ethics Seminar; Tulane University; The Ethics of Coordination Workshop at the Institute of Future Studies; Stockholm University; Utrecht University; Kerah Gordon-Solomon's seminar at Queen's, and the Bled Ethics Conference.

References

Cullity, G. 2003. "Asking Too Much." *The Monist*. https://www.jstor.org/stable/27903832.

Cullity, Garrett. 2004. *The Moral Demands of Affluence*. Clarendon Press.

Igneski, Violetta. 2006. "Perfect and Imperfect Duties to Aid." *Social Theory and Practice* 32 (3): 439–66.

Nefsky, J. 2017. "How You Can Help, without Making a Difference." *Philosophical Studies*. https://idp.springer.com/authorize/casa?redirect_uri=https://link.springer.com/article/10.1007/s11098-016-0808-y&casa_token=msrEC2TyzLsAAAAA:M7i5yQ3eh_UbFTfloJE5_h83iaIa3GaBafuStlj3SeLEOfFAIM_v1LpsNA5W4K9T4NYaymwv-y5Ih3NyOfM.

Noggle, Robert. 2009. "Give till It Hurts? Beneficence, Imperfect Duties, and a Moderate Response to the Aid Question." *Journal of Social Philosophy* 40 (1): 1–16.

Pummer, Theron. 2023. *The Rules of Rescue*. New York: Oxford University Press.

Schmidtz, D. 2000. "Islands in a Sea of Obligation: Limits of the Duty to Rescue." *Law & Phil.* https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/lwphil19&section=38&casa_token=_paYuZZg-XgAAAAA:2y6wnnX-L1fdzikDgh6M2X7AFw2ghY-9bVlfAPpURvVhyvAz25YzA24koT5agMWpMbcdphONIA.

need not have this structure. Probably just saying "thank you" will fully discharge my duty. But in such cases, it is not clear that gratitude has the relevant structure of imperfection.

Singer, Peter. 1972. "Famine, Affluence, and Morality." *Philosophy and Public Affairs* 1 (3): 229–43.———. 2019. *The Life You Can Save: How To Do Your Part To End World Poverty*. The Life You Can Save.org.

Stohr, Karen. 2011. "Kantian Beneficence and the Problem of Obligatory Aid." *Journal of Moral Philosophy* 8 (1): 45–67.

Temkin, Larry S. 2022. *Being Good in a World of Need*. Oxford University Press.

Tenenbaum, Sergio. 2024. "Can't Kant Count: Innumerate Views on Saving the Many over Saving the Few." In *Oxford Studies in Normative Ethics Volume 13*, edited by Mark Timmons, 215–34. New York: Oxford University Press.

Thomson, Jordan Arthur. 2021. "Relief from Rescue." *Philosophical Studies*, 1–19.

Timmerman, Travis. 2015. "Sometimes There Is Nothing Wrong with Letting a Child Drown." *Analysis* 75 (2): 204–12.

Woollard, F. 2015. *Doing and Allowing Harm*. Oxford: Oxford University Press.

Anne Schwenkenbecher[1]

Solving Collective Action Problems? We-reasoning as Moral Deliberation

Moral agents facing collective-action problems regularly encounter a conundrum: together, we can effect change whereas, individually, we are inefficacious. Further, what appears individually rational can be collectively suboptimal. An individual agent may employ different types of reasoning in deciding how to act vis-à-vis such problems. Reasoning in the I-mode, she takes her individual agency and efficacy in the world as the starting point: What is the best thing she can do given the circumstance and given what others do? It is act-based, best-response reasoning. The preferences of agents deliberating in the I-mode may well be other-regarding: e.g. they may aim at furthering the group's interest or collective good. We-mode reasoning, or 'we-reasoning', in contrast, is pattern-based: we infer our course of action from what is collectively best by way of acting as part of the group rather than for the sake of the group. I-mode reasoning with pro-group preferences (pro-group I-mode reasoning) and we-reasoning will often generate the same result, in particular in so-called strict joint necessity cases – where each agent's contribution is necessary for realizing a specific collectively available option. I-mode reasoning will regularly generate socially suboptimal results in so-called wide joint necessity cases – such as voting or carbon footprint reductions. Moral deliberating agents use both kinds of reasoning and contextual factors seem to function as important triggers. But can we-reasoning help us determine our moral obligations vis-à-vis collective action problems?

[1] Murdoch University, A.Schwenkenbecher@murdoch.edu.au.

1. Introduction

Environmental degradation and global climatic change are collective action problems. These problems are collectively caused and are only collectively solvable. More importantly, they generate rational and moral challenges and are, thus, often portrayed as dilemmas: what is individually optimal is collectively suboptimal. Famous examples of such dilemmas include the *tragedy of the commons* (ToC) and the *prisoners' dilemma* (PD).

Because of their unique structure, collective action problems regularly invite defection and free-riding. To the extent that the benefits of a collective good (as in the benefits of herd immunity achieved by high compliance with vaccination regimes, for instance) apply to all in a group – including those who failed to contribute to the production of the good – there exists an incentive to free-ride on others' contributions. Worse, still, in the prisoners' dilemma the best option for each player simply is the one where they defect while the other complies (even if it is a collectively suboptimal option) and there is the real danger of being the 'sucker' if one chooses to comply (wherein one gets made significantly worse off by the others' defection. In other words, in the PD (as well as ToC) there is a price to pay for complying (or cooperating, or contributing) while others defect. Further, there is the problem of individual inefficacy – no individual agent can unilaterally secure or undermine the collectively optimal outcome through defection – even those morally motivated vis-à-vis collective action problems may see this as grounds for not contributing. Ultimately, though, this approach to the collective action problem – "what should *I* do given the situation – what is *my* best response independently of others' choices?" – makes everyone worse off: in the standard solution to PD and the standard portrayal of ToC all end up with a scenario that is worse for them individually than if they had cooperated with the other player(s).

The standard solution to such problems is to change their incentive structure, their internal 'logic' if you will: The tragedy of the commons, for instance, is avoidable through governance (either through regulating the commons or turning them into private property). Environmental regulation may limit air and water pollution by making it preferable (individually rational) for the individual agent to choose a course of action that forms part of the optimal collective pattern.[2] For strategic interaction games, like the prisoners' dilemma, changing the incentive structure in experimental settings through repeating the game, for instance, will increase the frequency with which participants opt to 'cooperate' to produce the collectively optimal solution. More on this later.

[2] That is, if penalties for non-compliance are set at the right level and there is effective enforcement.

Standard solutions to collective action problems, then, make the collectively optimal choice individually optimal: through either increasing the cost of defection from the optimal collective pattern or lowering the potential cost for individual contributions to that pattern or both. Crucially, these solutions require external intervention – usually by an agent with the power to change the incentive structure. In the absence of such an agent, collective action problems tend to remain unresolved. Global climate change is a prime example of a (very complex) collective action problem and – in the absence of an agent with the above-described powers – the global climate regime has been failing to meet its most important goals such as limiting global warming to a maximum of 1.5° C.

Environmental degradation and climate change are also *moral* problems and they are moral problems of a special kind: our intuitive responses *as well as* our traditional moral theories[3] regularly fail to single out the morally optimal outcome where that outcome can only be collectively secured.[4] When faced with collective moral action problems, as individual deliberating agents we tend to feel torn between these two choices: (a) acting towards the collective good where the success in securing that good depends on others' compliance or contributions, and (b) unilaterally pursuing an individually achievable if morally suboptimal outcome (Schwenkenbecher 2021). Both our traditional theoretical repertoire and our intuitive responses make us prone to what Derek Parfit called 'mistakes in moral mathematics' (1984): our individual inefficacy in those cases makes us misjudge the moral status of our individual contributory actions both for positive (beneficial) collective actions and negative (harmful) ones. When we cannot unilaterally secure or prevent a collective outcome nor – as is often the case – make a perceptible difference to it, we tend to dismiss the idea of having moral obligations to contribute to the production or prevention of such outcomes.

This paper defends an alternative approach to thinking about these cases: 'we-reasoning' about our obligations vis-à-vis collective moral action problems (see also Schwenkenbecher 2019, 2021). My notion of 'we-reasoning' is based on Raimo Tuomela's pioneering work in philosophy of sociality wherein he posits the explanatory and normative importance of what he calls the 'we-mode' for understanding the social

[3] When I have used this term in the past, I have been asked what I mean by 'traditional moral theory'. This refers to (at the very least) the three best known groups of theories such as Virtue Ethics, Deontological Ethics and Consequentialist Ethics. See also my exchange with William McBride in Schwenkenbecher 2023 (*Social Philosophy Today*).

[4] Note: collective *moral* action problems are not those where people fail to produce a collective good because it is not in their self-interest to contribute, that is, because they act immorally. Rather, these are problems where even if each agent in a group is morally motivated, neither intuitive responses nor traditional moral theories will reliably point them towards the (set of) choice(s) that secures the collectively optimal outcome.

world (see, for instance, 1984, 2007, 2013)[5]. From there, the concept found its way into nonstandard game theory and the works of Michael Bacharach (2006) and into the wider philosophical discussion.[6]

We-reasoning – the way I use the term – constitutes a type of agency transformation in the way a collective action problem is approached by an individual deliberating agent (Schwenkenbecher 2019, 2021). It is reasoning in the we-mode as opposed to the I-mode. Instead of considering the problem from the point of view of the individual (what is the best thing *I* can do?), agents reason from the point of view of the group. They ask: what is the best thing *we* can do and – therefore – what is it that *I* need to do? (ibid).[7] I will explain this in more detail in a moment. But before doing so, we will need to introduce another conceptual distinction.

2. Joint Necessity Cases: Strict and Wide

Let us look at different collective action scenarios more closely. We can see that there are – very roughly – two types of scenarios:

(1) *Strict joint necessity cases* are those collective actions scenarios where the number of available agents (or contributors) equals the number of agents that are minimally necessary for realizing the collective outcome (or performing the collective action). Dancing tango is a type of collective action that requires at least (and at most) two people. Where two agents are present, each of them is needed to contribute *and* each agent is individually able to undermine the success of the collective action. No individual agent can dance tango by herself and whether or not she succeeds in dancing tango depends on the other person's ability and willingness to do so. In other words: in strict joint necessity scenarios *all* available agents must contribute to the joint endeavour in order for the collective outcome to be realised. *Each* individual agent has the power to unilaterally prevent the collective outcome, to *not* make it happen (Schwenkenbecher 2021: 8).

[5] Donald Regan (1980) worked on group-based reasoning even earlier than Tuomela.

[6] I cannot do justice to the entire literature around 'we-mode', 'we-reasoning', 'team-reasoning' and related concepts here, but will only point to some of the key authors: Sugden (2015), Gold & Sugden (2007), Hakli, Miller et al. (2010). Suffice it to say that Tuomela's work is much more broadly focused on sociality, in general, wheres Sugden's, Bacharach's and Gold's focus more narrowly on game theory and collective decision-making.

[7] Another way to put this is that in the I-mode agents are only able to select strategies whereas in the we-mode they can select outcomes (Hakli, Miller et al. 2010: 298).

'Typical' collective action problems have a different structure: they are *wide joint necessity cases*, and they are the ones I am most interested in.

(2) *Wide joint necessity cases* are those collective action scenarios where there are *more* available contributors than minimally necessary for realizing the collective outcome. Collective action problems are typically of this kind. The best example is vaccination against infectious diseases and the public good of herd immunity. In order for a group (e.g. the members of a political community) to achieve herd immunity against an infectious disease such as measles, not every member of the community has to be vaccinated against that disease. A 95% vaccination rate is deemed sufficient for generating herd immunity: the removal of the pathogen from that community and the resulting protection of all community members (vaccinated or not) from the disease. Unlike strict joint necessity cases, in these kind of scenarios no individual group member can unilaterally undermine the collective outcome. My not getting vaccinated (taken in isolation) does not jeopardize herd immunity. It is in jeopardy only if too many group members fail to contribute to the collective good. Voting in a referendum is another case in point: my vote is not going to make a difference to the outcome (or, more accurately, it is *extremely* unlikely to do so). My failure to vote in a referendum is not going to prevent (or produce) a desirable outcome.

It is in wide joint necessity cases that the so-called 'paradox' of collective action emerges: it may be collectively rational to jointly generate a certain outcome but it is individually rational to save oneself the effort of contributing and have others secure the collective good through their aggregate contributions. And so it is individually rational for each member of the group to do what is collectively not rational.

Morally speaking, the conundrum is this: If my failure to contribute to the production of a morally desirable collective outcome (e.g. a public good) is not making a difference to the outcome, then it appears that failing to contribute is not morally problematic. If this is true of one group member's failure to contribute, then it is true of every group member's failure to contribute. So, bizarrely, it would seem that no single group member has acted wrongly whenever a group of people fail to produce a morally desirable collective good (in wide joint necessity cases). In fact, for every individual contributory action to a morally desirable collective good we might find a competing individual action that directly and unilaterally secures an individually achievable goal that is also morally desirable. Hence, we end up with an analysis where we might at the same time condemn the collective failure to secure a particular good (for instance, herd immunity, or the legalization of abortion via a referendum)

but also grant that no individual had an obligation to contribute to securing that good.[8]

Derek Parfit's example of the 'harmless torturer' – the one who inflicts a very small (imperceptible) amount of pain onto each one of their thousand victims – is another case in point (1984). If a thousand torturers each inflict the same very small (imperceptible) amount of pain onto each one of their thousand victims, then there will be a thousand victims in a lot of pain – because the 'harmless torturers'' contributions add up. But – bizarrely and also wrongly, as Parfit explains – on individualist versions of consequentialism no one appears to be doing anything wrong. After all, what each person does – taken in isolation – is not harmful (as in painful) to any of the individual victims.

One way to move beyond this impasse is to move away from analysing these problems purely through an individualist lens. Parfit suggested that instead of focusing on individual acts and their effects, we should ask ourselves:

> Will my act be one of a set of acts that will *together* harm other people?' the answer may be Yes. And the harm to others may be great. If this is so, I may be acting *very* wrongly, like the Harmless Torturers. (1984: 86)

In other words, in order to assess some action's rightness or wrongness we must look at the outcome that we do (or could) produce *together with others* who are similarly placed. It is the collective level then that the wrongness (or rightness) of my individual action – and, therefore, its mandatory character – depends on (or is derived from). Collective action paradoxes – if they are paradoxes – disappear if we approach collective action problems that are wide joint necessity cases from the 'point of view of the group', that is, if we treat the collective level as primary.

3. We-reasoning Explained (In More Detail)

At this point, then, let us return to the notion of we-reasoning (or 'we-mode reasoning' for Tuomela) that was introduced earlier. It is an alternative method of reasoning about one's choices vis-à-vis joint necessity cases.

One way to describe we-reasoning is as top-down reasoning, starting from the most desirable collectively available outcome (something that can only be jointly

[8] Such a conclusion would be based on an assumption that is rarely if ever made explicit: If there exist two mutually exclusive courses of action and only one of them definitively makes a difference to whether or not a morally desirable good is secured then this course of action is morally superior to the alternative course of action. This is a moral difference-making principle, which, if interpreted individualistically (as it standardly is), privileges individually efficacious action over contributory action especially in wide joint necessity cases.

secured). Rather than choosing between individual options for action (or strategies), the we-reasoner chooses – if you will – between different (group-level) outcomes. The first step in the process of we-reasoning is what I call we-framing:

> *We-framing* means to include collectively available options in one's option set when deliberating about which option is best and identifying an option that is only collectively available as optimal. (Schwenkenbecher 2021: 13)[9]

This happens when I as a deliberating agent interpret (or perceive) a collective action scenario as a problem for 'us' – me and the other member(s) of my group. In practice it means that I will include options that are only collectively available in the set of options over which I am deliberating (that is, the set of options for acting that I am choosing from in my deliberation) (ibid., 2021). Those options concern outcomes that I cannot secure on my own – which is the characteristic feature of joint necessity cases.

The best way to illustrate this is by using a basic cooperative game: the Hi-Lo game:

Table 1: Payoff-matrix for Hi-Lo game

		Player 2	
		A	B
Player 1	A	Hi/Hi	0/0
	B	0/0	Lo/Lo

Hi > Lo > 0

In a Hi-Lo game, actual players tend to choose A over B. There are two different ways of making their choices at the individual level:

- If the other player chooses A then I am best off to also choose A, however, if the other player chooses B then I am best off to choose B.

Note that on this type of best response reasoning – or reasoning in the I-mode, the players will not arrive at a definitive conclusion or action recommendation, they end up with a conditional conclusion instead.

It should be noted that there is another way of reasoning in the I-mode, which avoids a conditional conclusion. Here, a player in the *I-mode* might reason that

[9] This is my definition of the term. For earlier and related uses see Bacharach (2006) and Butler (2012).

- If she chooses option A then she will either get the highest payoff if the other player also chooses A, or she'll get nothing if the other player chooses B, that is, if their choices do not match.

- On the other hand, if she chooses B, then she'll either get the lower payoff – if the other player also chooses B – or she'll get nothing if the other player chooses A, that is, if their choices do not match.

- Each player might then conclude that out of the set of possible outcomes [Hi or 0] and [Lo or 0] the first one is preferable because the lowest possible payoff is zero in both cases while the highest possible payoff is [Hi] if choosing option A [Hi]. Note that in this case, the players are not choosing an outcome as such but only a strategy that can lead to two possible outcomes.

In contrast, in the *we-mode*, a player will reason as follows:

- A/A [Hi/Hi] is the best possible outcome, therefore I should choose option A [Hi].

It is in that sense that the agent using we-mode reasoning (or we-reasoning) is choosing (group) outcomes and not (individual) strategies (Hakli et al. 2010: 298).

According to Hakli et al. (2010), only in the we-mode does the agent select an outcome as such, so only the we-mode guarantees that the best outcome (the Pareto optimal equilibrium in this case) is chosen. Work in experimental economics supports the assumption that people do in fact reason in the we-mode, at least sometimes (Butler et al. 2011; Butler 2012; Colman et al. 2008).

While the Hi-Lo game is an easy starting point for explaining we-reasoning, its real workings become more salient when we move to a competitive game like the Prisoners' Dilemma (PD). What is interesting about experimental evidence in relation to the PD is that players often do cooperate (Butler et al. 2011; Butler 2012) – in contrast to what conventional game theory predicts (or deems to be the rational choice). Let us have a closer look:

Table 2. Payoff-matrix for two-player PD game

		Player 2: Cooperate	Player 2: Defect
Player 1	Cooperate	R/R	S/T
	Defect	T/S	P/P
	T > R > P > S	R+R > T+S > P+P	

In the PD game, the highest payoff for an individual player is T (= temptation): it is part of an outcome that she can only achieve if she has opted to 'defect' while the other player chose to 'cooperate'. In other words, one player's achievement of the highest payoff requires the other player's ending up with the worst payoff. The lowest outcome for a player is S (= sucker) – where she cooperates while the other player defects.

The best 'group outcome' – the highest combination of payoffs – is R/R: the outcome where both cooperate.[10] However, each player in this game has an incentive not to cooperate: after all, if they defect then they can be made even better off – individually – than if they were to cooperate (as long as the other player cooperates, anyway). It is well known that the Prisoners' Dilemma is a scenario where each player's I-mode reasoning about their best individual choice ends up making both worse off: in their attempt to maximize their chance at receiving the highest individual payoff and to avoid being the 'sucker', each player chooses to 'defect' (this is the 'dominant strategy' in game-theoretic terminology) and both end up with the second worse outcome: P/P when they could have secured the – individually *and* collectively – better outcome R/R. The standard game-theoretic solution concept, the Nash Equilibrium, leads to this Pareto-inefficient outcome. It is an example of reasoning in the I-mode:

> In the *I-mode*, a player will reason that if she defects then she will either get the highest payoff – if the other player should choose to cooperate – or she'll get the second lowest payoff – if the other player should choose to defect as well.

[10] This may not be obvious from the payoff ordering in the table. However, in most formulations of the PD game that come with numerical payoffs, the combined payoff of R/R is greater than the combined payoffs for any of the other options. In that sense, the best 'combined' outcome is R/R whereas the best individual outcome is T.

Each rational player would conclude that out of the set of possible outcomes [T or P] for choosing to defect and [R or S] for choosing to cooperate the first one is preferable: The worst possible outcome when defecting – [P] – is still better than the worst possible outcome when cooperating [S] while the best possible outcome when defecting – [T] – is better than that of cooperating [R].

Note that both players reasoning in this way means that they will end up choosing the second *worst* individual and combined outcome [P/P] when they could have secured the preferable second *best* outcome [R/R]. In other words, each individual choosing the better set of possible payoffs guarantees the worse outcome in this case – both at the group level and the individual level.

The Prisoners' Dilemma is regularly considered to reflect the underlying structure of many social action problems, including environmental challenges, with its payoff-structure (or incentive structure) to closely resemble that of problems such as environmental pollution, global warming caused by greenhouse gas emissions, and – generally – the degradation of common or shared resources, for instance.[11] According to this interpretation, agents in those kinds of scenarios if acting rational will choose to 'defect' – that is, to not contribute towards environmental goals such as reducing pollution or to undermine collective goods by actions such as overstocking the commons or overfishing shared fish stocks. It is important to note that this is not an empirical claim about how (and why) all (or most) agents in these situations *do* act.[12] But rather it is a way of explaining the emergence of collective action problems and an attempt at understanding them from the point of view of the individual agent.[13]

However, as it turns out, in real life people sometimes choose to cooperate in the Prisoners' Dilemma and not everyone will overexploit common goods and resources even if they could. The standard explanation of cooperative behaviour in PDs in experimental settings has been to suggest that players are not fully rational or that their preferences may be group-regarding or other-regarding (which suggests that a payoff transformation has occurred – this essentially means that we are no longer looking at a PD since the change in preferences means a change in payoffs and payoff structure).

[11] E.g. Gardiner, 2006. It should be noted that this interpretation is not universally shared. See FN 12.

[12] In fact, many people *do* try to reduce their individual carbon footprint, for instance, with a view to contributing to the collective goal of reducing greenhouse gas emissions (and potentially mitigating climate change).

[13] There might be good reason to be cautious with this kind of interpretation. Matthew Kopec has argued that the PD interpretation of the international climate regime deadlock, e.g., could also be a self-fulfilling prediction (2017). Aklin and Mildenberger argued that there is no empirical evidence supporting the view that climate change policy is a "global collective action problem structured by free-riding concerns" (2020: 4, see also comment by Kennard and Schnakenberg 2023).

An alternative explanation of cooperative behaviour in PD games and of players choosing the Pareto optimal equilibrium in Hi-Lo games has been suggested by advocates of we-reasoning (Tuomela 1984, 2007, 2013, Bacharach 2006, Gold & Sugden 2007, Hakli et al. 2010, Butler 2012): players may be employing something other than individual-based best-response reasoning in 'solving' these collective action puzzles. They may be engaging in we-reasoning or team-reasoning where they identify the cooperative solution [R/R] to be the best overall outcome therefore choosing to play their part in securing that outcome, namely to cooperate.

Jurgis Karpus and Natalie Gold put it as follows:

> The key difference here is that individualistic reasoning is based on evaluating and choosing a particular strategy based on the associated expected personal payoff, whereas team reasoning is based on evaluating the outcomes of the game from the perspective of the team, and then choosing a strategy that is associated with the optimal outcome for the team. (2017: 402)

Susan Hurley writes:

> Participating in collective activity rather than acting as an individual can be instrumentally rational, by reference to the ends of a component of the relevant collective. (2005b: 594)

Empirical studies in experimental economics have provided *some* evidence to believe that this is how some players arrive at their decision to cooperate in a PD (Butler 2012; Butler et al. 2011, Colman et al. 2008, Karpus & Gold 2017)[14]. What is more, proponents of we-reasoning (Bacharach 2006, Gold & Sugden 2007, Hakli et al. 2010) suggest that it is *rational* to reason this way, opposing standard game theory's notion of rationality and rational choice.

Scholars who write on we-reasoning or team-reasoning disagree on when (and why) people team-reason, including whether or not the choice of frame is itself a rational or even conscious choice. Certain features of the decision scenario are thought to increase or decrease the likelihood of agents' framing the decision problem as a problem for her individually or as a problem for the group. *(i) Strong interdependence*, according to Bacharach will increase the likelihood of we-framing (2006, see also

[14] Karpus' and Gold's discussion includes an important caveat though: "There is a major difficulty that any empirical test of team reasoning will unavoidably face: the fact that a number of separate hypotheses are being tested at once.... Also, if decision-makers do not follow individualistic best-response reasoning in certain situations, we need to be able to distinguish team reasoning from other possible modes of reasoning that they may choose to endorse" (2017: 407).

Karpus & Gold 2017). Strong interdependence occurs "when there is a Nash equilibrium that is worse than some other outcome in the game from every player's individual point of view." (ibid., p. 403) as is the case for both the Hi-Lo game and the Prisoners' Dilemma. The latter, however, also displays the *(ii) double crossing feature* – "the possibility of an individual personally benefiting from a unilateral deviation from the team reasoning solution" (Ibid.). According to Bacharach (2006), this will reduce the likelihood of we-framing (See also Smerilli 2012). Another aspect that may impact on an agent's framing of a decision problem is that of *(iii) group identification*: whether or not the agent perceives herself as belonging to the same social group as the other player(s) (Bacharach 2006). Further, Colman et al. discuss *(iv) risk dominance* as inhibiting the choice of optimal collective options – where the latter also come with the risk of players receiving the lowest payoff (2008: 395).

4. We-reasoning As a Moral Deliberation Strategy

We-reasoning is a rational deliberation strategy for joint necessity cases. In abstract games or vignettes, the value of different outcomes is expressed in terms of payoffs. Higher payoffs for the individual means a better outcome for that player or agent. A higher combination of payoffs signifies a better outcome for the group or combination of agents. Sometimes, as in the Hi-Lo game the best outcome for the group will correspond to the highest possible payoff for each individual. In the PD game it does not. Here, the player who chooses to defect is better off than the one who does not – unless both players defect. In any case, both 'conventional' game theory with its individualistic best-response reasoning and we-reasoning (or team-reasoning) are about rational choices, not about moral choices. Karpus and Gold argue that

> Taking goals as given to us by our theory of value, or moral theory, turns team reasoning from a theory of rational choice into a theory of moral choice, which is not intended by many of its proponents. (2017:405)[15]

Yet, exploring we-reasoning in moral deliberation is precisely what I do in this article and have done in some of my previous work on this topic (Schwenkenbecher 2019, 2021). Moral collective action problems do regularly have the same structural features as strategic interaction scenarios. The payoffs for each player and the outcome for the group (or set of players) in strategic interaction scenarios can refer to anything that

[15] A previous attempt at combining the two was made by Donald Regan (1980) – however, it focused only on utilitarian ethics whereas my theory is largely neutral with regard to the substantive moral theory (for a discussion of theory neutrality see Schwenkenbecher 2023).

the agent(s) consider(s) valuable or in their interest. In experimental settings, players are usually offered money.

Moral deliberation is the activity by which we determine what is the morally right thing to do. In moral deliberation we choose the morally best course of action in a given situation, weighing up different courses of action, where each would have some (positive or negative) moral value attached to it.[16] Sometimes, the morally best outcome we can produce is one where we need to cooperate or at least coordinate with other agents. Those are the kinds of cases I am interested in here – collective *moral* action problems.

In many such scenarios, morally valuable outcomes can be produced by agents pursuing individually available options. This means that individually available options for action compete with collectively available options. Take a scenario with the structure of a stag hunt game as an example:

Table 3. Payoff-matrix for two-player stag hunt game

		Player 2: Stag	Player 2: Hare
Player 1	Stag	10/10	0/3
	Hare	3/0	3/3

Each player in this game must choose between hunting stag or hunting hare. They can only successfully hunt stag together, whereas they can successfully hunt hare on their own. If one player chooses the 'stag' strategy and the other player does so as well both players receive the maximum individual payoff *and* achieve the best group outcome. If players 'do their own thing' and hunt for hares they still benefit, but significantly less. The worst scenario is being the only one choosing the cooperative strategy, i.e., to hunt stag. Importantly, the cooperative choice (hunt stag) competes with the non-cooperative choice (hare) in that both convey some benefit (where the latter is risk-dominant over the former). Options for moral action can be structured a similar way. Where they do, choosing to contribute to what is overall morally optimal competes against the best outcome individuals can produce unilaterally or independently of others' choices (that is, we-mode reasoning competes against best response

[16] 'Moral value' is used ecumenically here: it can refer to the best outcome or the right type of action (see Schwenkenbecher 2023).

reasoning). In large-scale wide joint necessity cases, there is an added complication: I-mode reasoning about our moral choices appears to come at no moral cost because the contributions individuals can make to improving or worsening large-scale collective action problems are so minute and seem morally negligible (see also Parfit 1984).

In any case, it is fair to assume that if we switch between modes of reasoning in strategic interaction cases then we probably do the same in *moral* deliberation. But let us take a step back and look at simpler, small-scale joint necessity scenarios to illustrate how we-reasoning happens in moral deliberation. *Rescuers:* Imagine a rescue scenario wherein a drowning person can only be saved by two agents acting in conjunction. Garrett Cullity (2004) describes a version of this scenario where two people have to jointly operate a winch to get another person to safety. Let us assume the alternative course of action has them call emergency services or go look for a lifeguard. This alternative course of action is morally worse than operating the winch because it comes with a significantly lower probability of saving the drowning person.

Table 4. Moral value-matrix for two-person rescue case (Rescuers)

		Beach goer 2	
		Operate winch	Find lifeguard
Beach goer 1	Operate winch	10/10	0/3
	Find lifeguard	3/0	3/3

My assumption here is that the overwhelming majority of agents when facing a scenario like *Rescuers* (i.e. with this kind of structure and transparency concerning the moral values of the outcomes) will pick the cooperative strategy ('operate winch'). And my contention is that they *ought to* pick it despite the fact that the morally optimal outcome is not individually available, but only collectively available (more on that later). In picking this outcome they (potentially) we-frame the scenario and – therewith – include a collectively available option in the set of options for action over which they deliberate.

Whether or not moral deliberators do in fact we-reason is, of course, an empirical question. But it is no less probable for moral deliberators to engage in we-reasoning than it is for deliberators in non-moral strategic interaction. In either case, we may infer that agency transformation and we-reasoning form part of the best explanation for said choices.

This argument should become stronger once we look at some real-world cases of collective moral action. Let me begin with a scenario observed at a train station in Perth, Western Australia, in August 2014:

> *Commuters:* On a busy weekday morning a man gets trapped between a commuter train and the station's platform. He will be crushed should the train move. Dozens of people who happen to be on the platform witnessing his predicament join forces in pushing the train to tilt it away from the man. Together they manage to free him, therewith saving his life. (Schwenkenbecher 2019, 153)

Dozens of commuters push against the train in order to free the trapped man. Best-response reasoning about the morally right course of action is unlikely to have prompted that kind of response: each commuter had reason to assume that their contribution was unlikely to make a difference to whether or not the desirable outcome would be achieved. I-mode (or best-response) reasoning would have produced, at most, a conditional obligation: "I should contribute to this joint endeavour if I make a positive difference to the optimal outcome. Whether or not I make a difference depends on (a) how many people it takes to tilt the train and on (b) how many are contributing already." Not only does this conditional obligation depend on facts unknown and possibly unknowable to the agent (in the situation). What is more: if everyone only has a conditional obligation where does that leave people in the group? It leaves them without any clear answer as to whether or not they should contribute.

Worse, still: in the I-mode it would appear that people could easily reason their way out of an obligation to contribute: Assuming everyone has some morally relevant goals competing with that of pushing the train[17] (such as arriving at work on time or honouring whatever time-sensitive commitments commuters tend to have on a weekday morning), each individual agent might plausibly reason: "In not contributing I am very unlikely to undermine this collective endeavour. In other words, the success of this endeavour does not (or is very unlikely to) positively depend on my contribution. Therefore, I should pursue an alternative course of action where I am very likely to make a positive difference to a morally desirable outcome, namely continuing on my way to whatever commitment I have already made and am keen to honour." When reasoning in the I-mode about their obligations, each commuter is

[17] One might be tempted to compare such competing goals to the 'double-crossing feature' of some games: whereas in a PD, e.g., unilateral deviation from the 'cooperative' strategy benefits an individual, one could argue that in some collective moral action cases unilateral deviation may generate a greater overall benefit. This would be the case where the group outcome would be secured independently of the deviating agent's contribution, that is, in cases where the outcome is overdetermined (wide joint necessity cases). However, the 'double crossing feature', does not map very well onto the structure of those cases (Pareto-optimal outcomes in multi-player PD games are not overdetermined).

justified in concluding that they need not contribute. Thus, no one has an obligation to help push against the train.

And, this could indeed be all there is to say about this kind of scenario if it were not for two issues: the empirical fact that enough people *did* contribute. And the fact that we tend to see this collectively available outcome as the morally best course of action. The second point is one about moral intuitions, for what they are worth: I think we agree that people *ought to* have helped the trapped commuter. (See also Schwenkenbecher 2019, 2021).

These two issues might prompt us to look for alternative solution concepts as well as an alternative explanation of observed behaviour: we-framing the scenario as a problem for the group and then enacting the strategy that corresponds to the optimal (collective) outcome will get the commuters to reliably choose to push the train. Also, it is – arguably – a better explanation of the observed behaviour, not least because it is a simpler explanation (Ibid.).

Let me bring up two more examples before we get to address some important caveats and limitations: We are regularly being encouraged to reduce our individual carbon footprint through behavioural change (e.g. as consumers) with a view to contributing to a global effort to reduce greenhouse gas emissions and mitigate climate change. I take it to be a fact that many people not only do reduce (or at least are mindful of) their individual carbon footprint but that they do so – at least in part – because they think it is the right thing to do. That it is the 'right thing to do' is unlikely to be based on the assumption that they – individually – are making an actual, measurable difference to the desired goal. Individually, they are not 'difference-makers' in any morally significant sense. It is more plausible to think of such everyday contributions to mitigating climate change as examples of people enacting their part in what they perceive to be an optimal or at least morally valuable large-scale pattern of action.[18] In other words: it is an individual *playing her part* in producing a morally desirable *collective* outcome.[19] She derives her individual course of action from that collective goal.

This is speculative, of course.[20] Further, I do not pretend to suggest that this is the only or even the dominant motivation for people who change their behaviour to be more 'environmentally conscious'. My main contention is that such considerations

[18] Though, arguably, most people might be ignoring the most impactful course of climate action they could individually take: having fewer children (Wynes et al. 2017).

[19] See also Christopher Woodard (2003).

[20] There does not seem to be any empirical literature on why people reduce their carbon footprint or act more sustainably. However, there exists research in social psychology into the motivating factors for people contributing to collective endeavours such as donating to charity for poverty relief (Thomas 2009a, 2009b, 2010, Thomas et al. 2009). Social psychologists show that collective capacity – the idea of being part of a group and of making a difference as part of that group – plays a key motivating role. This supports my argument here.

make sense – both from a rational and from a moral point of view.[21] It is a plausible way to conduct moral deliberation vis-à-vis large-scale collective action cases (with wide joint necessity). By that I mean that – in principle – it is at least on a par with I-mode reasoning in those cases.

Let us have a look at another example. Vaccinations usually gain their efficacy from two sources: individual immunity (active or passive, that is, through either the production of antibodies against a pathogen or through the direct injection with antibodies) plus herd (or collective) immunity. Herd immunity is achieved when rates of individual immunity are high enough for the pathogen to disappear from a population. For measles the vaccination rate to achieve herd immunity is roughly 95% of the population. If the vaccination rate drops then herd immunity is lost (as was the case in France in 2010-2011). Especially during the COVID-19 pandemic, the argument from the *collective* benefits of widespread vaccination played a major role in public health campaigns. What is more, this type of message clearly resonated with people. For instance, Joshua Lake et al. "found that the message expressing self-transcendence values was ranked most persuasive by 77% of respondents" in the Australian context, e.g. (2021). While there was an individual benefit to be had, many Australians seem to have acted also for the collective benefit of getting vaccinated. They chose to play their part in what they perceived to be the collectively optimal pattern of action.[22]

To conclude, I have invited my readers to consider the possibility that some of our moral reasoning is (or resembles) we-reasoning by providing examples where we-reasoning provides a very good explanation for observed choices. However, I have not actually delivered a decisive argument in favour of the claim that (i) people *do* in fact we-reason in moral deliberation and even less of an argument for the claim that (ii) people *ought to* we-reason in moral deliberation, at least some of the time. These questions must be left for another paper.

[21] The underlying assumptions here is, of course, that such behavioural change *does* make sense, e.g. that carbon footprint reductions *are* a good idea.

[22] We-reasoning may also explain what is often referred to as the 'voting paradox': people vote despite not being difference-makers in elections. Since voting is somewhat costly, the question is why they bother? (Obviously, this question does not arise where voting is compulsory, in countries such as Australia, e.g.). We-reasoning provides an explanation and a solution – if you will – to the paradox: voters are playing their part in what they perceive to be a worthwhile collective endeavour (regardless of whether they are making an actual difference to the outcome) (see also Hurley on Quattrone's and Tversky's voting experiment, Hurley 2005a: 204-5).

References

Aklin, M. and M. Mildenberger (2020). Prisoners of the Wrong Dilemma: Why Distributive Conflict, Not Collective Action, Characterizes the Politics of Climate Change. *Global Environmental Politics* 20(4), 4-27.

Bacharach, M. (2006). Beyond individual choice: Teams and frames in game theory. Princeton, Princeton University Press.

Butler, D. (2012). A choice for 'me' or for 'us'? Using we-reasoning to predict cooperation and coordination in games. *Theory and Decision* 73, 53—76.

Butler, D. J., V. K. Burbank and J. S. Chisholm (2011). The frames behind the games: Player's perceptions of prisoners dilemma, chicken, dictator, and ultimatum games. *The Journal of Socio-Economics* 40, 103—114.

Colman, A.M., Pulford B.D., Rose J. (2008). Collective rationality in interactive decisions: Evidence for team reasoning. *Acta Psychologica* 128(2), 387—97.

Cullity, G. (2004). The moral demands of affluence. Oxford, Clarendon Press.

Gardiner, S. M. (2006). A perfect moral storm: Climate change, intergenerational ethics and the problem of moral corruption. *Environmental Values* 15, 397—413.

Gold, N. and R. Sugden. (2007). Theories of team agency. *Rationality and Commitment*. F. Peter and H. B. Schmid, Oup Oxford: 280—312.

Hakli, R., K. Miller and R. Tuomela (2010). Two kinds of we-reasoning. *Economics and Philosophy* 26, 291—320.

Hurley, S. (2005a). Rational agency, cooperation and mind-reading. In Gold, N. (Ed.), *Teamwork* (pp. 200—215). Palgrave Macmillan, London.

Hurley, S. (2005b). Social heuristics that make us smarter. *Philosophical Psychology* 18, 585—612.

Karpus, J. and N. Gold (2017). Team reasoning: Theory and evidence. Kiverstein, J. (Ed.), *The Routledge Handbook of Philosophy of the Social Mind*n (pp. 400—417). Routledge.

Kennard, A. and K. E. Schnakenberg (2023). Comment: Global Climate Policy and Collective Action. *Global Environmental Politics* 23(1), 133-144.

Kopec, M. (2017). Game theory and the self-fulfilling climate tragedy. *Environmental Values* 26(2), 203-221.

Lake, J., P. Gerrans, J. Sneddon, K. Attwell, L. C. Botterill, and J. A. Lee. (2021). We're all in this together, but for different reasons: Social values and social actions that affect Covid-19 preventative behaviors. *Personality and Individual Differences* 178, 110868.

Parfit, D. (1984). Reasons and Persons. Oxford, Oxford University Press.

Regan, D. (1980). Utilitarianism and Co-Operation. Oxford University Press.

Tuomela, R. (1984). A Theory of Social Action. Dordrecht, Springer Netherlands.

Tuomela, R. (2007). The Philosophy of Sociality: The Shared Point of View. New York, Oxford University Press.

Tuomela, R. (2013). Social Ontology: Collective Intentionality and Group Agents. New York, Oxford University Press.

Schwenkenbecher, A. (2019). Collective moral obligations: 'we-reasoning' and the perspective of the deliberating agent. *The Monist* 102(2), 151-171.

Schwenkenbecher, A. (2021). Getting our act together: a theory of collective moral obligations. New York, Routledge.

Schwenkenbecher, A. (2023). Commentary for NASSP Award Symposium: Response to Commentators. *Social Philosophy Today* 39: 215-226.

Smerilli, A. (2012). We-thinking and vacillation between frames: filling a gap in Bacharach's theory. *Theory and Decision* 73(4), 539-560.

Sugden, R. (2003). The logic of team reasoning. *Philosophical Explorations* 6, 165—181.

Sugden, R. (2015). Team Reasoning and Intentional Cooperation for Mutual Benefit. *Journal of Social Ontology* 1(1), 143–166.

Thomas, E. (2010). Social psychology of making poverty history: Motivating anti-poverty action in Australia. *Australian Psychologist* 45, 4—15.

Thomas, E. F. (2009a). Transforming "apathy into movement": the role of prosocial emotions in motivating action for social change. *Personality and Social Psychology Review* 13, 310—333.

Thomas, E. F. (2009b). The role of efficacy and moral outrage norms in creating the potential for international development activism through group-based interaction. *British Journal of Social Psychology* 48, 115-134.

Thomas, E. F., C. McGarty and K. I. Mavor (2009). Aligning identities, emotions, and beliefs to create commitment to sustainable social and political action. *Personality and Social Psychology Review* 13, 194—218.

Woodard, C. (2003). Group-based reasons for action. *Ethical Theory and Moral Practice* 6, 215—229.

Wynes, Seth, and Kimberly A. Nicholas. (2017). The climate mitigation gap: Education and government recommendations miss the most effective individual actions. *Environmental Research Letters* 12, 074024.

Olle Torpman[1]

Responsibility-Based Reasons to Act in Collective Impact Cases[2]

What moral reasons to act could an individual have if her action would not make any difference? In this paper, I argue that there are responsibility-based reasons for individuals to act, and that these can help explain why an individual sometimes should act in so-called collective impact cases even if she cannot make a difference with respect to the outcome in those cases. I distinguish between retrospective and prospective kinds of responsibility, and argue that (i) an individual has prospective responsibility-based reasons to act in a specific way in collective impact cases, if she will thereby avoid contributing causally to harm, or contribute causally to good when that is desirable; and (ii) an individual has retrospective responsibility-based reason to act in a specific way in collective impact cases, if she would otherwise be blameworthy for making a (causal or constitutive) contribution to harmful outcomes in such cases.

[1] Institute for Futures Studies, correspondence: olle.torpman@iffs.se.

[2] Funding from Riksbankens Jubileumsfond (grant number: P22-0662) is gratefully acknowledged.

1. Introduction

It is common to think that an individual has no reason to act specifically in a situation if she cannot make any relevant difference to the outcome in that situation. For example, if it does not matter to climate change whether or not you stop emitting, then you have no reason – at least no climate change-related reason – to stop emitting. This poses a problem particularly in so-called *collective impact cases* where the impact (good or bad) stems from individuals' collective actions but where no individual member of the collective seems to be capable of making a relevant difference to that impact. Intuitively, however, it seems that individuals should at least sometimes act specifically even in such cases. But what moral reasons to act specifically could an individual have in collective impact cases if her action will not make any difference to the occurrence of the outcome?

There are basically two possible strategies of arguing for the existence of reasons for an individual to act specifically in such cases. One strategy would be to object to the idea that an individual's action in collective impact cases does not make any relevant difference, and show *how* it actually can make such a difference. In this regard, some have argued for the existence of expected utility-based reasons, according to which an individual has reason to act in virtue of the chance (however small) that her action will pass a threshold that leads to a (much) better outcome (see, e.g., Kagan 2011). Voting cases seem to be such cases, where there is initially a small chance for each vote that it will make a difference as to who will win.

The other strategy would be to argue that there are other reasons besides those that are *difference-based*, as we may call them, and that such other-based reasons are present in collective impact cases. In this regard, some have argued for the existence of *fairness-based* reasons, arguing that the collective has a duty to act in collective impact cases, and that the only fair thing to do for an individual member of this collective is to participate in the work that is needed (e.g., Cullity 2000; Baatz 2014). Others have argued in favor of *virtue-based* reasons for actions, the idea being that we have reason to perform actions that stems from virtuous motives or character traits, whether or not these actions make any difference to the outcome (see, e.g., Jamieson 2007; Hourdequin 2010). Another proposal refers to *helping-based* reasons, the idea being that an individual act can help to bring about an outcome in the sense that it makes a non-superfluous causal contribution to that outcome, even if it cannot make a difference to the outcome, and that this can in itself have reason-giving force (see Nefsky 2017). Yet others have argued in favor of *participation-based* reasons, where an individual has reason to participate in group activities that can make a difference, even if the individual herself cannot make this difference (e.g., Wieland & Oeveren 2020).

My aim in this paper is to offer another proposal along the lines of the second strategy, which, I will argue, is less problematic than the existing proposals of that kind. More precisely, I will argue for the existence of *responsibility-based* reasons as a distinguished type of reasons for individuals to act specifically in collective impact cases. In section 2, I briefly clarify the different notions of responsibility that are relevant to the present paper, distinguishing between prospective moral responsibility and retrospective moral responsibility. In section 3, I discuss prospective moral responsibility-based reasons for action. In section 4, I discuss retrospective moral responsibility-based reasons. Section 5–8 answers potential objections to the responsibility-based reasons account, most of which have been raised against other accounts in the debate. Section 9 concludes.

2. Different types of responsibilities

There are many different types of responsibility discussed in the philosophical literature (see, e.g., Williams 2010; Poel, Royakkers, & Zwart 2015). For instance, an agent can be *causally* responsible for something in the sense that she caused it, or *attributively* responsible in the sense that it is attributable to her agency, or *morally* responsible in the sense that she is either praise- or blameworthy for it or under a duty to do something about it. This paper is concerned with moral responsibility. More precisely, I will investigate the reasons for action that moral responsibility can ground in collective impact cases.

As the above description unveils, there are (at least) two types of moral responsibility. On the one hand, an agent is *retrospectively* morally responsible if and only if she is worthy of praise or blame for her choices of actions, or the outcomes of her actions. Roughly speaking, retrospective moral responsibility is backwards-looking, and regards *things one has done (or omitted doing)*. On the other hand, an agent is *prospectively* morally responsible if and only if she has a certain duty to act with respect to a certain situation – e.g., to care for someone, to solve a problem, or to pay certain costs. Prospective moral responsibility is thus forward-looking, and means responsibility *to do something* or, in other words, *to see to it that something is the* case (Poel, Royakkers, & Zwart 2015).

Any account of responsibility identifies what we may call *responsible-making features* – that is, the (set of) features in virtue of which an agent is responsible. Different responsible-making features might be relevant depending on whether we have retrospective moral responsibility or prospective moral responsibility in mind. One such feature that is relevant to retrospective responsibility concerns foreseeability, in the sense that an agent can be worthy of blame or praise for an action or outcome only if she understands the situation and can foresee the connection between her

action and the outcome. Another such feature concerns voluntariness, implying that an agent can be praised or blamed for an action only to the extent it is within her own control. A third retrospective responsible-making feature concern avoidability, in the sense that an agent can be blamed for a choice of action only if she could have chosen otherwise. One of the prospective responsible-making features concerns ability, in the sense that an agent can be responsible to do something only if she has the ability to do so. I will get back to these differences below.

While neither causal nor attributive responsibility in themselves implies reasons for action, moral responsibility might do. Given the two types of moral responsibility clarified above, there are potentially two different types of moral responsibility-based reasons: (i) retrospective responsibility-based reasons, and (ii) prospective responsibility-based reasons. Below, I will discuss both types. Given the tight connection between prospective responsibility and duties, it is clearer that prospective responsibility might yield reasons for action. Hence, I will start with that.

However, the connection between prospective responsibility and duties might put into question the relevance of the notion of prospective responsibility, and hence the relevance of prospective responsibility-based reasons for action. Why not just say that I have a duty or obligation to care for my daughter, and a duty or obligation not to do harm, etcetera, and skip the talk about prospective responsibility? Saying that an agent A has a prospective responsibility to Φ might thus be a different way of saying that A has a duty or obligation to Φ. If so, talk about prospective responsibility would be redundant and uninformative. Hence, it would make no sense either to talk about prospective responsibility-based reasons for individuals to act specifically in collective impact cases.

The way to address this worry, I think, is to point out that there is a sense of prospective responsibility that differs from duty and obligation. There are several ways in which this can be done. For instance, in the entry "Collective Responsibility" on the *Stanford Encyclopedia of Philosophy*, Marion Smiley says that "[i]n cases where we use the language of moral obligation, we signal that the agent has to perform a particular act. In cases where we use the language of responsibility, we allow the agent to use its own judgment in deciding how to bring about the desired state of affairs" (Smiley 2023). If this is correct, prospective responsibilities concern generic types of actions, whereas duties mainly concern sub-types and particular tokens of actions. This means that I might have a responsibility to do something in general, without having a duty to do anything in particular.

A similar view is found in Robert Goodin, who moreover argues that what matters to prospective responsibility is that the agent "*see[s] to it* that X" (see Goodin 1995: 83). He says that "'[s]eeing to it that X' requires, minimally; that [the agent] satisfy himself that there is some process (mechanism or activity) at work whereby X will be

brought about; that [he] check from time to time to make sure that that process is still at work, and is performing as expected...” (ibid.).

This idea is shared by Ibo van de Poel, Lambèr Royakkers, and Sjoerd D. Zwart, who moreover argue that the “...sense in which responsibilities are different from duties [...] is that responsibilities do not require the agent to achieve the outcome φ by her own actions” (Poel, Royakkers, & Zwart 2015: 28–29). The idea is, in other words, that prospective responsibilities can, whereas duties cannot, be fulfilled by external factors. They can, for instance, be delegated to other agents or realized by natural causes. Again, what matters is that the agent *sees to it that* φ. This moreover implies, they argue, that “[a]lthough this responsibility is aimed at realizing φ, the occurrence of φ is not the main criterion whether [an agent] actually fulfilled her forward-looking responsibility” (Poel, Royakkers, & Zwart 2015: 29). Interestingly, this moreover means that an agent can fulfil her prospective responsibility to see to it that φ even in cases where φ does not occur.

It is also possible that prospective responsibilities can *ground* duties, meaning that they would be more fundamental than duties. Overall, on the basis of these observations, I will assume that there is a notion of prospective moral responsibility that is not redundant, and, hence, that it makes sense to investigate the possibility of prospectively moral responsibility-based reasons for actions.

3. Prospective responsibility-based reasons: Do no harm

Prospective moral responsibility connects agents with possible future actions. As mentioned above, it provides reasons for seeing to it that a certain state of affairs obtains. As I see it, prospective moral responsibility can yield reasons to act in basically two ways: (i) to abstain from wrongdoing in the first place (i.e., unconditionally); and (ii) to correct for wrongdoings that have already taken place (i.e., conditionally). Since most collective impact cases discussed in the literature do not involve prior wrongdoing, I will here focus on unconditional prospective responsibility. Note that I here use ‘wrongdoing’ in a broad sense to be compatible with different moral theories.

I assume there are two types of unconditional wrongdoings in this regard: to contribute causally to the presence of harms where avoidable, or to contribute causally to the absence of benefits where desirable. I here use “harms” and “benefits” also in a broad sense to be compatible with different moral theories. A necessary condition for wrongdoing is, thus, to contribute causally to harms where avoidable, or to not contribute causally to benefits where desirable. This implies that a sufficient condition for abstaining from doing wrong is to contribute causally to neither the presence of avoidable harms, nor the absence of desirable benefits. In fact, I can see no other way

in which an agent can abstain from wrongdoing. If I am correct about that, then this is also necessary for abstaining from wrongdoing.

Consequently, individuals have prospective moral responsibility-based reasons to not contribute causally to harm where avoidable, and to contribute causally to benefits where desirable, in collective impact cases. This means that the relevance of prospective moral responsibility-based reasons for individuals to act specifically in collective impact cases thus hinges on the meaning of "causal contribution". There are different ways in which "causal contribution" could be analyzed. For the sake of argument, I will here assume the so-called NESS (Necessary Element of a Sufficient Set) account, which takes a cause to be a necessary element of a set that is sufficient in the circumstances for their effects (see, e.g., Beebee & Kaiserman 2020). This view builds on the views of J. L. Mackie (1965) and has in different versions been proposed by, for instance, Braham and van Hees (2009) and Kaiserman (2016). The definition provided by the NESS account can be formulated as follows:

> **An agent A (in circumstances C) contributes causally to an outcome O if, and only if, A performs an action such that (i) the action is a member of a set of actions that is sufficient (in C) for O, and (ii) no subset of that set of actions is sufficient (in C) for O.**

This merits clarification. First, it implies that the set of actions is minimally sufficient for O, meaning that that there is no proper subset relative to the set at issue that would also realize O. This does not require that there is no proper subset relative to the full set of involved actions. For example, if you and four other agents act in a way that leads to O, but only four agents' actions are necessary for the realization of O, then there are several subsets of actions – e.g., the original set of actions minus your action – that would also have realized the outcome. This, however, does not mean that your action does not contribute causally to the outcome. Given that you actually perform your action in this case, your action is itself a member of a set (indeed four sets) which is minimally sufficient for the realization of the outcome. This means that you do make a causal contribution in this case. This moreover shows that the definition applies to cases of overdetermination.

Second, it is important to note that the occurrence of "performs an action" is a simplification. In fact, what is relevant is what the agent chooses to do – whether it is to act or to omit. If intentional, an omission could also make a causal contribution, since an agent's choice to omit can constitute a member of a set that is sufficient for the realization of an outcome. Suppose that it is enough that three out of four people intentionally omit to push a button in order to realize O. If all four intentionally omit to push the button, then each of their individual intentional omissions belongs to a

set that is minimally sufficient for the realization of O. This moreover implies that the only way in which an agent can make sure not to contribute to the outcome in such a case is to push the button. When doing so, the agent's choice no longer belongs to any set of actions that is minimally sufficient for that outcome. Still, the definition has the welcome implication that it does not count all omissions as causal contributions to outcomes. Indeed, omissions sometimes do not constitute members of any set of actions that are minimally sufficient for realizing the outcome at issue. If you choose to omit to take part in a joint activity that saves 10 lives, and if your omission has no effects on that activity, then your choice does not belong to any set of acts that is minimally sufficient for realizing the outcome. Hence, your omission does not count as a causal contribution to saving those lives.

Equipped with this notion of causal contribution, we can draw some conclusions regarding the prospective moral responsibility-based reasons for individuals to act specifically in collective impact cases. In general, it gives an individual reasons to (i) abstain from taking part in any collective activity that produces harm where avoidable, and to (ii) take part in collective activities which produce benefits where desirable, since by doing so she sees to it that her action (i) is no member of any set that is minimally sufficient for the presence of that harm, and (ii) is a member of a set ot actions that produce that benefit. In the case of climate change in particular, the NESS account of causal contribution implies that an individual can fulfill her prospective unconditional responsibility by not emitting. Only thus can she see to it that her action is no member of any set that is minimally sufficient for the production of harmful climate change.

Of course, objections may be raised. But since many objections apply equally well to the account of retrospective responsibility-based reasons, I will first have a look at that account.

4. Retrospective responsibility-based reasons: Avoid blame

There is a widespread view in the literature on moral responsibility that an agent can be retrospectively morally responsible for an outcome if and only if they voluntarily, foreseeably, and avoidably contribute somehow to that outcome (see, e.g., Williams 2010; Braham & van Hees 2012; Poel, Royakkers, & Zwart 2015; Goodin 2018). This means that contribution, voluntariness, foreseeability, and avoidability are conditions for retrospective moral responsibility. Consequently, if an agent A fulfills these conditions with respect to a certain outcome O, then A is retrospectively morally responsible for O. In addition, if O is (sufficiently) morally bad or undesirable, then A is

blameworthy for O. If O is instead (sufficiently) morally good or desirable, then A is praiseworthy for O.

Under the plausible assumption that agents have reason, other things being equal, to avoid blame, we may assume that they have reason, other things being equal, to end any relevant responsibility-relation between themselves and the outcomes for which they are blameworthy. Given the conditions for retrospective moral responsibility, this means that the agent has a reason to avoid contributing voluntarily and knowingly to such outcomes.

Although it is possible in principle to fulfill this requirement through involuntary or ignorant action, it is hard to see how this would be possible in practice. If an agent puts himself in a situation where he is forced to do something (in order to fail with respect to voluntariness), then he would most certainly be blameworthy for having put himself in that situation. Likewise, if an agent puts himself in a state of ignorance (in order to fail with respect to foreseeability), then he would most certainly be blameworthy for having put himself in that state. This indicates that the only practical way of avoiding blame for a certain outcome is to not contribute to that outcome.

This line of reasoning suggests that there are retrospective moral responsibility-based reasons – or blame-avoidance reasons – for action. Moreover, it suggests that individual agents thus have such reasons not to contribute to morally bad or undesirable outcomes – be them individually or collectively produced.

Note that I have so far left it open what type of 'contribution' is at stake in the case of retrospective moral responsibility. The reason is that some have argued that causal contribution is not necessary for retrospective responsibility. Braham and van Hees, for instance, argue that "...holding a person morally responsible in the sense of blameworthiness appears to require something weaker than actual causal contribution to some state of affairs. A person may be blameworthy if, inter alia, the action they performed is at least a potential causal factor" (Braham & Hees 2009: 342). If this is correct, there is an interesting difference between the conditions for prospective and retrospective moral responsibility, respectively.

Robert Goodin (2018) offers another suggestion along these lines. He says that even if you do not causally contribute to a certain outcome stemming from a collective activity, you could still be constitutively responsible for it in virtue of taking part in, and hence being part of, the collective activity as a whole which produces this outcome. Being constitutively responsible in this respect, he argues, simply means being a part of a whole. This means that human beings and their actions can be constitutively responsible for group activities of which they are part.

Moreover, Goodin argues that "[y]ou bear constitutive responsibility in the sense that you are liable to credit or blame for voluntarily and knowingly being a part of that whole" (2018: 41). Presumably, if the outcomes of such activities are (sufficiently)

bad, and if the agent contributes voluntarily and knowingly, then that agent is blameworthy for taking – and hence being – part of that activity. If the outcome is (sufficiently) good, the agent is instead praiseworthy. Given that agents have blame-avoidance reasons for action, and given that constitutive responsibility in collective harm cases implies blameworthiness, an agent has retrospective responsibility-based reasons not to take part in such activities. This holds whether or not it could be argued that they contribute causally to that outcome.

Summing up thus far: An individual has prospective responsibility-based reasons to act in a specific way in collective impact cases given that she will thereby avoid contributing causally to harms where avoidable, or contribute causally to benefits where desirable. Also, an individual has retrospective responsibility-based reason to act specifically in collective impact cases given that she will otherwise be blameworthy for making a voluntary, foreseeable and avoidable (causal or constitutive) contribution to harmful outcomes in such cases.

Let us now consider objections.

5. First objection: The problem of emergent properties

The above proposed account(s) of responsibility-based reasons for individuals to act specifically in collective impact cases could be objected to by pointing out a distinction between aggregative and emergent properties. This argument has been raised by Kingston & Sinnot-Armstrong (2018). Although they raise it against what they call "the partial causation argument" (i.e., against the applicability of the notion of causal contribution) in collective impact cases, it might also apply to the notion of constitutive responsibility, hence making it potentially effective against both types of responsibility-based reasons for action.

Aggregative properties are properties that belong to both parts and wholes, where the property of the whole equals the aggregate sum of that same type of property of the parts. Size and weight are examples of such properties. If every single piece of the puzzle is 3 cm^2, then the thousand-piece puzzle as a whole is 3000 cm^2. Or, to use Kingston & Sinnot-Armstrongs example: "[C]onsider a quantity of oil that has a mass of one kilogram and contains, say, 3 times 1025 molecules of oil. Then we can calculate the mass of one molecule of oil simply by dividing one kilogram by 3 times 1025" (Kingston & Sinnott-Armstrong 2018: 175).

Emergent properties, on the other hand, belong only to the whole, and are hence not properties of its parts. Such properties emerge out of parts that lack that property. Kingston & Sinnot-Armstrong exemplify:

> The quart of oil is very slimy, but an individual molecule of oil by itself is not slimy at all. It is not that the molecule has a little sliminess, but much less than the whole quart. An individual molecule is not slimy in the least. We cannot feel any individual molecule at all, so it cannot feel like slime. The same point applies to other properties of the oil, including appearing yellowish and causing moving parts to last longer. (2018: 175)

On their view, climate change is emergent in this way. They say that, "[j]ust as individual molecules of oil do not cause parts of sensations of sliminess (or yellowish color), so individual molecules of greenhouse gas do not cause parts of dangerous climate impacts. Instead, as with the sliminess and color of oil, what increases the dangerous impacts of climate change is larger groups of molecules of greenhouse gases" (2018: 175). What is more important, is that they take the emergent property of climate change to imply that individual emitting actions, lacking that property, cannot be partial causes of climate change. Against their opponents in the debate about "joyguzzling" (joyriding in a gas guzzler) as an example of questionable emitting activity, they say that if "...global climate change as well as specific climate events and their harms are all emergent phenomena [...] they cannot cite partial causation to justify their claim that there is a moral requirement to refrain from joyguzzling" (2018: 176).

However, the mere distinction between aggregative and emergent properties does not rule out that emitting actions can be parts (i.e., members) of wholes (i.e., sets) that are themselves minimally sufficient for climate change – yet climate change is an emergent property which is lacking in individual emitting actions. If an individual emitting act is such a member, then it is a causal contributor, whether or not climate change is an emergent property. Consider voting for example. No single vote for candidate A possesses a 'winner-making' feature. But if more than 50% of the electorate vote for A, then these votes will together possess that feature. This means that 'winner-making' is an emergent property. Nevertheless, individual votes may contribute causally to A's election win. Consequently, just because climate change would be emergent rather than aggregative, this does not imply that emitting activities could not contribute causally to climate change. Kingston & Sinnot-Armstrong are therefore wrong when they say that "the partial causation argument [...] assumes that climate change is aggregative, not emergent" (2018: 178).

Still, their objection might have force against the applicability of the notion of constitutive contribution in the context of retrospective responsibility-based reasons. It seems plausible, for instance, to say that one molecule of oil is not a constitutive part of the sliminess of a gallon of oil. Likewise, it seems plausible to say that a single act of emissions is not a constitutive part of the climate change harm. The underlying explanation would be that no individual act (such as an emitting action) which lack

an emergent property (such as climate change harm) can constitute a part of an emergent property (such as climate change harm).

At a closer look, however, this seems to be false. Suppose that I knowingly and voluntarily add a certain chemical, C1, into a bowl, another person knowingly and voluntarily adds another chemical, C2, into that same bowl, a third person knowingly and voluntarily adds yet another chemical, C3, into that bowl, and these three chemicals together give rise to a chemical composition with a corrosive emergent property that is lacking in each of the single chemicals, C1-C3, and in each of the three pairs of them. Even if it cannot be said that our individual actions (of adding a single chemical into a bowl) are constitutive parts of the corrosiveness as an emergent property of that chemical composition, we might say that our individual actions are constitutive of the chemical composition as such. This shows that a non-emergent action could be a constitutive part of a whole that gives rise to an emergent property. I might thus be constitutively responsible for that. If someone is harmed by the corrosive chemical composition, for example, I would be blameworthy for contributing constitutively to its cause. The same seems to hold in the climate case: Even if my individual emissions would not be constitutive of any climate change harm as such, I may be blameworthy for acting in a way that is constitutive of the 'cloud' of emissions, as it were, that causes climate change harm.

6. Second objection: Non-threshold cases

There is a certain type of collective impact cases that appears to pose a problem for a responsibility-based account of reasons for action. These are called "non-threshold cases" (Nefsky 2017) or "non-triggering cases" (Tiefensee 2022), which are distinguished from so-called "threshold cases" or "triggering cases". Nefsky explains the difference as follows:

> In threshold cases, for each outcome of the morally significant sort in question, there is some precise number of acts of the relevant type needed to bring it about: any less will not be enough to bring it about, and any more will not change things with respect to that outcome. If a threshold is hit exactly, though—as in the case of a tie or a one-vote-win—then each act can make a difference. In non-threshold cases, on the other hand, there is no precise number of acts of the relevant kind needed for a given outcome. While acts of a certain type together cause (or are part of what cause) the outcome in question, there is no sharp boundary between enough such acts and not enough. So, in non-threshold cases you cannot have enough acts for a particular outcome without having more than enough such acts, and thus taking one away will never itself make a difference. (2017: 2746)

Tiefensee exemplifies this difference by saying that "whereas election wins are clear examples of triggering phenomena, in that victory is generally secured upon reaching the precise threshold of 50% of the votes plus one, no such precise threshold appears to exist in relation to air pollution [or] water contamination..." (2022: 3308).

The possibility of non-triggering cases assumes vagueness in the form of semantic or metaphysical indeterminacy. In the case of climate change this means, Tiefensee points out, that "while some amount of greenhouse gases is sufficient for global warming to be harmful, which exact amount this is remains vague" (2022: 3311). More precisely, she thinks that non-triggering cases requires metaphysical indeterminacy. In the climate case, this means that there is no fact of the matter as to which precise amount of greenhouse gases would be minimally sufficient to bring the collective harm of global warming about (see also Kingston & Sinnot-Armstrong 2018).

As Tiefensee mentions, however, this possibility hinges on a number of controversial assumptions. First, it assumes that there is in fact metaphysical indeterminacy. Second, it assumes that climate change is of such kind. Being aware of these controversies, she emphasizes that she will not commit herself to these assumptions, but rather investigate what would follow if they were true (2022: 3309).

She discusses two different interpretations of metaphysical indeterminacy in this respect. On the first, "what is metaphysically indeterminate is *when* the increasing amounts of CO2 molecules become sufficient to cause harmful global warming" (2022: 3311-2, my emphasis). There is thus no precise threshold after which, but rather a range or interval within which, the relevant climatic effect may be caused. On the second interpretation, "metaphysical indeterminacy could be understood along the lines of *ontic* indeterminacy. More precisely, we could argue that the cloud itself is an ontically indeterminate object, such that there is simply no fact of the matter as to which molecules are part of it" (2022: 3321, my emphasis).

The problem with non-triggering cases, involving metaphysical indeterminacy of either of these types, is that the standard notions of causal and constitutive contribution, respectively, appear inapplicable. In such cases, an individual agent's choice of action seems not to constitute any member of any set of actions that is minimally sufficient for the realization of the collective impact, and might not even be a constitutive part of the whole that causes it. If climate change harm is non-triggering, it would be hard to explain how individuals are contributing (causally or constitutively) to it (see, e.g., Wieland & van Oeveren 2020: 175-6). Since causal and constitutive contribution is a condition for prospective and retrospective moral responsibility, respectively, it seems that we thus have to accept that individuals lack any responsibility-based reasons to act in a specific way in non-triggering collective impact cases.

I think this conclusion is too hasty, however. If metaphysical indeterminacy is real in the sense that there is no fact of the matter as to which precise amount of emissions

is minimally sufficient to bring about the collective harm of global warming, or which emissions end up as constitutive parts of the 'cloud' as a whole which causes this warming, then this plausibly means that it is indeterminate as well whether or not a specific agent's action will be a member of any such set, or a constitutive part of such a whole. However, this does not imply that our emitting actions are *never* members of sets of actions that are minimally sufficient for the realization of harmful climate change. Nor does it imply that our individual emitting actions *never* end up as constitutive parts of the 'cloud' that in effect causes such harm. Rather, it implies that our emitting actions *sometimes* are members of such sets, as well as constitutive parts of such wholes. As Tiefensee puts it:

> [A]t the moment of releasing CO_2 molecules, we do not know where these molecules will end up: Will they remain totally detached from the cloud, such that they have nothing to do with the cause of harmful global warming? Will they find themselves in the cloud's centre, such that they determinately belong to the cause of this collective harm? Or will they end up in the shaded areas, such that there is no fact of the matter as to whether or not the molecules we release are part of the harmful cloud, and thus part of the cause of the collective harm? (2022: 3322)

This suggests that metaphysical indeterminacy – if real – implies epistemic uncertainty: If there is indeterminacy in the world, then we cannot know if or when (or which of) our actions belong to which of these categories. Even if climate change (or any other collective impact case) is non-triggering, an individual's action *might* end up in a set of actions that is minimally sufficient for the realization of the undesired outcome, or become a constitutive part of the whole which causes this outcome. Although we will not be able to determine exactly *which* emitters make such contributions, we are able to establish that non-emitters certainly *do not* make such contributions.

Other things being equal, it is plausible to assume that an individual is prospectively responsible to not risk making contributions to harm. And this gives her reasons not to take such a risk. Since it can moreover be argued that taking such risks is blameworthy, she would in addition have retrospective responsibility-based reasons pointing in the same direction. Hence, individuals would have both prospective and retrospective moral responsibility-based reasons to not take part in collective harm cases – whether or not they involve thresholds or metaphysical indeterminacy. In the climate case, the only way in which the agent can make sure her emissions do not end up in the 'cloud' of emissions that causes harmful climate change, is to not emit.

7. Third objection: The problem of overriding reasons

The account of responsibility-based reasons might appear to yield too strong reasons for individuals to act specifically in collective impact cases. To see this, suppose that a construction worker is about to fall down from the top of a wobbly scaffolding, unless all of the five and only bystanders step in to stabilize it. You are one of these five bystanders. As you happen to know, however, none of the other bystanders will step in. Hence, you know that your decision to step in can make no difference with respect to the construction worker eventually falling. In this case, it might seem implausible to say that you should step in.

Nevertheless, since the only way in which you can abstain from contributing causally (or constitutively to the cause of) the fall of the construction worker is to step in, the account I have proposed implies that you do have a responsibility-based reason to step in. Indeed, that is the only way in which you can see to it that your action will not constitute a member of a set of actions that is minimally sufficient for the construction worker's falling, or not become a constitutive part of the cause of that fall. Do we hence have a reductio argument against the account of responsibility-based reasons?

No. What explains the intuition that you should not to step in in the wobbly scaffolding case, is not that you cannot make any difference by stepping in, but rather that it makes a difference in some other respect not to step in. In most real-world cases like this, stepping in would cost time and involve risks to oneself – which could be avoided by choosing not to step in. The mere fact that an action cannot make a difference in some respect can never in itself be a reason not to perform it unless there is some alternative action the performance of which can make a difference in some (perhaps other) respect.

We hence need to distinguish between pro tanto reasons and all-things-considered reasons, of which only the former may be overridden by other more weighty reasons (see, e.g., Wieland & van Oeveren 2020). In the wobbly scaffolding case, the responsibility-based reason you have to step in is a pro tanto reason that is overridden by the reasons you have to not step in. In a situation where the other four bystanders would have stepped in, however, the high moral value of saving the construction worker from falling would imply that the reason for you to step in overrides the pro tanto reason (regarding costs of time and risks to yourself) to not step in. The lesson to learn from this is that just because one should not do X does not mean that one has no reason to do X.

But what if your only reasons to step in are responsibility-based reasons, and where

all other reasons – e.g., difference-based, and self-interested, etc. – point against stepping in? If they are weighty enough, I think we should just bite the bullet and accept that you have all-things-considered reasons not to step in. But what if it holds for all collective impact cases – say, that all other reasons together carry heavier weight and jointly recommend something different than the responsibility-based reasons? Then, of course, we would have to accept that our responsibility-based reasons for action would yield no concrete normative implications in such cases.

However, I do not think that is the case. First of all, the main reasons against stepping in (or in other ways acting in a specific way) in collective impact cases seem to be self-interested reasons, since doing so often requires a personal sacrifice. But it is not set in stone that such reasons always carry heavier weight than responsibility-based reasons in such cases. Moreover, it is not clear that an agent will always have self-interested reasons not to step in in collective impact cases. Sometimes she will benefit more from stepping in than from not. For instance, there are well known co-benefits from eating vegetarian instead of meat, as well as from taking the bike instead of the car to work, and so on.

Second, what an agent is morally required to do in cases of collective impact (as in any other type of case) is what she has all-things-considered reasons to do. And what she has all-things-considered reasons to do is determined by the weighing together of all pro tanto reasons she has in that situation. In the introduction, I briefly mentioned some such reasons for participating in collective impact cases – such as expected utility-based reasons, fairness-based reasons, virtue-based reasons, and helping-based reasons. Even if none of these pro tanto reasons would in isolation be capable of yielding any moral requirement of individuals to act specifically in collective impact cases, they might together be able to yield such a requirement.

8. Fourth objection: The problem of non-generalizability

In her criticism of other accounts of reasons for individuals to act in collective impact cases, Nefsky appears to implicitly assume what Andrea Asker (2023: 2384) explicates in a number of "success conditions" for such accounts. First and foremost, Asker explicates a "Generalizability condition", according to which "[t]he successful account should identify a weighty enough moral reason in all the collective impact cases of concern". As this means, an account of an individual's reasons to act specifically in collective harm cases should have something interesting to say in such cases. This condition seems to be implicitly assumed also by others in the debate (see, e.g., Kingston & Sinnot-Armstrong 2018).

The discussion in the previous section suggests that the account of responsibility-based reasons fails to meet the generalizability condition. For instance, if responsibility-based reasons are in some cases insufficient to generate moral requirements, perhaps due to the existence of overriding reasons, then it will not be able to "identify a weighty enough moral reason" in such cases. And if some collective impact cases involve metaphysical indeterminacy, and if the notions of causal or contributive responsibility does not apply in all of these cases, it means that the account of responsibility-based reasons might not apply to those cases either.

As Asker points out, however, it is not obvious that generalizability should be accepted as a condition for accounts of individuals' reasons to act specifically in collective impact cases. As she says, the best approach might well be "...a pluralistic approach, one that employs different accounts to identify moral reasons for individual action in different types of collective impact cases..." (Asker 2023: 2395). My previous arguments point in the same direction. If what an individual should do in a certain situation is a matter of what all-things-considered reasons for action she has in that situation, and that responsibility-based reasons constitute one type of pro tanto reasons that together with other pro tanto reasons determine her all-things-considered reasons, then it is simply implausible to assume that only one type of reason should identify a weighty enough moral reason in all the collective impact cases of concern. While responsibility-based reasons might be most salient in some collective harm cases, virtue-based, fairness-based, expected utility-based, helping-based – or any other relevant – reasons might be more salient in other such cases.

We should therefore accept that there might be cases – e.g., some non-threshold cases where an individual actually does not contribute (neither causally, nor constitutively) – where there are no responsibility-based reasons to step in. We should also accept that, if the individual also lacks any other-based reasons to step in or if she has stronger reasons not to step in – then she actually should not do so. This also suggests that it would be a mistake to assume from the start that all collective impact cases are such that they involve weighty enough reasons for individuals to step in.

8. Conclusion

In this paper, I have argued that an individual has two types of responsibility-based reasons to act in a specific way in collective impact cases: (i) she has prospective responsibility-based reasons to act, if she will thereby not contribute causally to the presence of harm where avoidable or to the absence of good where desirable; and (ii) she has retrospective responsibility-based reason to act, if she will otherwise be blameworthy for making a (causal or constitutive) contribution to harmful outcomes in such cases.

The responsibility-based account has the advantage not only of avoiding some of the problems to which other accounts are vulnerable, but also to answer some of the remaining issues observed by others in the debate. For instance, Tiefensee argues that, due to the possibility of non-triggering cases and metaphysical indeterminacy,

> [w]e must be able to show that individual agents have a reason to act in a specific way in view of a morally relevant aggregate effect E, even though their actions make *no difference* to E and they are *uncertain* whether or not there is a *fact of the matter* as to whether or not their actions are *partial causes of* E. (2022: 3322)

The account of responsibility-based reasons does just that. Moreover, Wieland and van Oeveren (2020: 185), defending the account of participation-based reasons, say that one remaining question related to their account is this: "why is participation morally significant [...]?" The account of responsibility-based reasons answers this question as follows: Participation is morally significant because it lets the agent fulfill prospective and/or retrospective responsibilities.[3]

References

Asker, Andrea S (2023) 'The Problem of Collective Impact: Why Helping Doesn't Do the Trick', *Philosophical Studies* **180**: 2377–97. doi:10.1007/s11098-023-01995-7.

Baatz, Christian (2014) 'Climate Change and Individual Duties to Reduce GHG Emissions', *Ethics, Policy & Environment* **17**: 1–19. doi:10.1080/21550085.2014.885406.

Beebee, Helen & Alex Kaiserman (2020) 'Causal Contribution in War', *Journal of Applied Philosophy* **37**: 364–77. doi:10.1111/japp.12341.

Braham, Matthew & Martin Hees (2009) 'Degrees of Causation', *Erkenntnis* **71**: 323–44. doi:10.1007/s10670-009-9184-8.

Braham, Matthew & Martin van Hees (2012) 'An Anatomy of Moral Responsibility', *Mind* **121**: 601–34.

Cullity, Garrett (2000) 'Pooled Beneficence', in Michael J Almeida, ed., *Imperceptible Harms and Benefits*: 1–23. Springer Netherlands.

Goodin, Robert E (1995) *Utilitarianism as a Public Philosophy*, Cambridge Studies in Philosophy and Public Policy. Cambridge University Press.

[3] I would like to thank Andrea Asker, Krister Bykvist, Daniel Ramöller, and the audience at the *Second Internal Workshop* on May 6 2024 at IFFS, for helpful comments.

Goodin, Robert E (2018) 'Constitutive Responsibility: Taking Part, Being Part', *Analysis* **78**: 40–45. doi:10.1093/analys/anx146.

Hourdequin, Marion (2010) 'Climate, Collective Action and Individual Ethical Obligations', *Environmental Values* **19**: 443–64. doi:10.3197/096327110x531552.

Jamieson, Dale (2007) 'When Utilitarians Should Be Virtue Theorists', *Utilitas* **19**: 160. doi:10.1017/s0953820807002452.

Kagan, Shelly (2011) 'Do I Make a Difference?', *Philosophy & Public Affairs* **39**: 105–41. doi:10.1111/j.1088-4963.2011.01203.x.

Kingston, Ewan & Walter Sinnott-Armstrong (2018) 'What?S Wrong with Joyguzzling?', *Ethical Theory and Moral Practice* **21**: 169–86. doi:10.1007/s10677-017-9859-1.

Mackie, J L (1965) 'Causes and Conditions', *American Philosophical Quarterly* **2**: 245–64.

Nefsky, Julia (2017) 'How You Can Help, without Making a Difference', *Philosophical Studies* **174**: 2743–67. doi:10.1007/s11098-016-0808-y.

Poel, Ibo van de, Lambèr Royakkers, & Sjoerd D Zwart (2015) *Moral Responsibility and the Problem of Many Hands*. Routledge.

Smiley, Marion (2023) 'Collective Responsibility', in Edward N Zalta and Uri Nodelman, eds., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2023/entries/collective-responsibility/.

Tiefensee, Christine (2022) 'Indeterminacy and Collective Harms', *Philosophical Studies* **179**: 3307–24. doi:10.1007/s11098-022-01827-0.

Wieland, Jan Willem & Rutger van Oeveren (2020) 'Participation and Superfluity', *Journal of Moral Philosophy* **17**: 163–87. doi:10.1163/17455243-01702002.

Williams, Garrath (2010) 'Moral Responsibility', *Oxford Bibliographies Online*.

IFFS Working Paper Series

The Institute for Futures Studies publishes preprints from research projects based at the institute. Find all of them at iffs.se/en. Below a list of our latest working papers.

2023:11 Scanlonian Contractualism and Future Generations
by Emil Andersson, Gustaf Arrhenius & Tim Campbell

2023:10 Do We Owe the Past a Future? Reply to Finneron-Burns
by Patrick Kaczmarek & S.J. Beard

2023:9 Population, Existence, and Incommensurability
by Melinda A. Roberts

2023:8 Longtermism and Neutrality about More Lives
by Katie Steele

2023:7 The Ethical Risks of an Intergenerational World Climate Bank (as Opposed to a Climate Justice World Bank)
by Stephen M. Gardiner

2023:6 Inducement-Based Emissions Accounting
by Olle Torpman

2023:5 Sex Selection for Daughters: Demographic Consequences of Female-Biased Sex Ratios
by Karim Jebari & Martin Kolk

2023:4 Uncertainty Attitudes as Values in Science
by Joe Roussos

2023:3 DALYs and the Minimally Good Life
by Tim Campbell

2023:2 How to Value a Person's Life
by John Broome

2023:1 How to Feel About Climate Change? An Analysis of the Normativity of Climate Emotions
by Julia Mosquera & Kirsti Jylhä

2022:13 Becoming a Business Student: Negotiating Identity and Social Contacts During the First Three Months of an Elite Business Education
by Anna Tyllström, Nils Gustafsson & Gergei Farkas

2022:12 Mot ett våldstabu? Våldsbrott i fransk lagstiftning från 1200-tal till 1800-tal
by Maria Wallenberg Bondesson

2022:11 More, Better or Different?
Gustaf Arrhenius & Klas Markström