Small Impacts and Imperceptible Effects: Causing Harm With Others*

Kai Spiekermann

Forthcoming in *Midwest Studies in Philosophy* XXXVIII (2014), ed. Peter French

1 Introduction

In an increasingly crowded and interactive world, there are more and more ways to harm people in an indirect way. These "new harms" (Lichtenberg, 2010, 558) are typically caused by many hands (Thompson, 1980). Many people use too many plastic bags, drive their cars too much, eat too much meat or bluefin tuna, drink bottled water, et cetera. Each individual act has a negligible effect, and may, as a singular act, not be harmful—but the same act performed by millions is. This gap between the (almost or perhaps entirely) harmless singular act and the harmful performance of the same act by many spells trouble for the moral evaluation of these acts and for assigning responsibility.

In a recent article, Shelly Kagan (2011) tries to dissolve some of these difficulties. In essence, Kagan claims that many small contributions must always encounter a threshold such that a relevant harm is triggered. If that is so, Kagan argues, any single action contributes to expected harm because there is a non-zero probability that it crosses the relevant threshold. Julia Nefsky (2011) pokes holes in Kagan's argument, showing that Kagan proceeds too quickly in dismissing the challenge of sorites-like situations where each additional action does not add additional perceived harm. This exchange shows that the issue of imperceptible effects and harms is still not

* Ideas for this paper were presented at the MANCEPT workshop on forward-looking collective obligations, at the NELPP seminar at the University of Newcastle, the Department of Government at the University of Essex, and the Collective Obligation Workshop in Manchester. I would like to thank all audiences for their comments and suggestions. I am particularly grateful to Christian List and Felix Pinkert for generous written comments. All errors are my own.

conclusively settled thirty years after Derek Parfit's (1984) seminal discussion in *Reasons and Persons*.

I propose to take the imperceptibility challenge seriously: our perception can be *minimal change insensitive* so that very small changes are impossible to perceive. And if the normatively relevant consequences of an action, holding all other actions fixed, cannot be perceived, then we are challenged to explain what makes the action wrong. My tentative answer is that an action cannot only be wrong because of its immediate effects, it can also be wrong because it can possibly cause an effect together with other actions. To avoid such wrong actions, an agent has to engage in forward-looking considerations as to how her actions, together with other actions that are possibly performed, can be harmful.

The paper proceeds by specifying the decision situations of interest. Section 3 states four prima facie attractive propositions about problems of imperceptible effects and shows that these four propositions are mutually inconsistent. Sections 4–7 discuss four attempts to avoid the inconsistency by relaxing different propositions. I sketch my preferred solution in section 8 and conclude in section 9.

2 Zooming in on the Problem

When many people make a small contribution to a large problem, we can distinguish between two schematic ways how these contributions lead to harm. Either the contributions are harmless until the overall contribution reaches a tipping point at which great harm is caused. Or the contributions gradually increase the level of harm, without any critical thresholds. In this paper, the focus is exclusively on the latter case.

Many hands problems are particularly challenging when the individual consequences of actions cannot easily be associated with harm because the changes brought about by each individual are so miniscule that they do not register as harm (I will flesh out this thought in much greater detail below). Some collectively caused harms may appear to have that structure, but, under closer inspection, we can rule out some of them. For instance, jointly causing climate change is often seen as a paradigmatic example (see Sinnott-Armstrong, 2005). When looking at the implications of individual acts of emissions, however, it turns out that they are often quite significant. The typical personal lifetime emissions in a developed country are

expected to "wipe out more than six months of a healthy human life" in the calculations of John Broome (2012, ch. 5). Smaller emissions also have clear effects. Flying from London to New York causes about one tonne of CO_2 emissions. The damage from these emissions can be quantified as a double digit US-$ amount[1], so they are likely to register. This shows that many larger emissions do cause expected harm, and this harm can be quite significant indeed.[2]

Sinnott-Armstrong and others have in mind smaller emission amounts: the emission of a Sunday pleasure drive with your SUV, for example. If the emission is quite small (perhaps a very short Sunday pleasure drive) it may not register on the scale of harm. These are the cases I am interested in here. Nevertheless, carbon emissions are not the best example to use: even though single instances of small emissions may only cause imperceptible effects, a different way to individuate actions would make them components of larger actions that do trigger perceptible effects (Andreou, 2006). So perhaps consequentialists have difficulties criticizing a single short pleasure drive, but they have the tools to criticize an individual's total annual emission pattern. In the case of emissions, that is the correct approach in the first place – what matters are aggregate emissions, after all.

The problem of many hands becomes much harder and more interesting if the individual actions (with imperceptible effects) are not performed repeatedly by the same individual. Consider a simplified version of Parfit's "Harmless Torturers" (Parfit, 1984, p. 80). A group of n torturers can each push a button to increase the voltage of the electric shock the victim[3] receives by the amount 1000 volt / n (one "notch"). The pain caused by 1000 volt is excruciating. It is plausible to assume that there is a number of torturers n so large that the victim cannot distinguish between x torturers and $x + 1$ torturers for any x between 0 and $n - 1$. Suppose the torturers push their button only once. If the harm done is defined as the additional suffering caused, then, it can be argued, none of the single torturers causes any harm. And,

[1] In estimated carbon prices deemed necessary for effective mitigation action.

[2] There may be objections against using *expected* rather than *actual* harm to morally evaluate these actions. It is true that, since the effect of the emissions may not linearly lead to more harm, one cannot claim that each individual emission will certainly lead to actual harm. However, increasing the probability of serious harm occuring in the future is surely morally wrong as well.

[3] In Parfit's original example the torturers have 1000 victims, but this is not necessary for my discussion.

unlike the case of carbon emissions, there is no larger individual action pattern that would cause any perceptible effects and could be criticized on the basis of the harm caused, since the torturers will not repeat their action.

3 The Contradiction

It is useful to make this setting more precise. Let there be a set N of individuals labelled $1, 2, \ldots, n$. Each individual can perform one of two actions, "contribute" or "not contribute". We denote a contribution by individual i with $a_i = 1$, and no contribution by $a_i = 0$. There are no other actions available to the individuals, which means that for all $i \in N : a_i \in \{0, 1\}$. We call the vector of all actions $\boldsymbol{a} = (a_1, a_2, \ldots, a_n)$ an action profile, while the sum of all contributions is $c = \sum \boldsymbol{a}$. For simplicity, assume that the individual actions are exchangeable, so that it does not matter who contributes, it only matters how many individuals contribute. The contributions make an impact m_c, and m is strictly increasing in c, such that each additional contribution increases the impact somewhat.

The impact can be measured precisely: the voltage, the poison emitted, the temperature increased by greenhouse gas emissions, et cetera. Therefore, one might think that the harm produced can simply be measured directly in terms of the impact. But this view is mistaken. The impact as such is morally neutral—what is normatively relevant is how the impact determines a morally relevant property such as harm. What is needed is a bridge principle connecting empirical facts with normative propositions. Hedonists, for instance, take pleasure and pain as morally relevant and say that an impact leads to harm if it increases pain, relative to a baseline. Of course, many other such principles are conceivable and have been discussed, but, for simplicity, I stick with the hedonistic line here. The important upshot is that the impact is *causally related* to the harm but it is not identical with the harm. Rather, the harm is a function of the impact, $h(m)$. I assume that the harm is weakly increasing in the impact (that is, more impact will either lead to equal or more harm).

While we can often quantify impacts precisely, this may not be true of the harm caused by impacts. Think about Parfit's harmless torturers again. Even though there will be a difference in voltage (i. e. in impact) when comparing x and $x + 1$ torturers,

there may be no difference in the perceived pain. And since Parfit proposes, at least initially, to measure harm in perceived pain, the torturers are a problem.

More generally, the problem of imperceptible effects arises as a contradiction between four (at least prima facie) plausible propositions:

Direct Act Consequences. Whether an action is wrong only depends on the consequences in a particular sense: the consequences of this particular action while holding all other relevant actions fixed.

Experienced Harm. For an action to be wrong (expected) experienced harm must be among the consequences.

No Stepwise Harm. There exist decision situations such that the individuals decide on their respective actions $a_1, a_2, \ldots, a_n$ and:

(i) if all individuals "contribute" ($c = n$) then the actions cause experienced harm, and if all individuals do "not contribute" ($c = 0$) the actions do not cause experienced harm.

Also, for any level of contribution $x \in (0, n-1)$:

(ii) adding one more contribution, so that $\sum \boldsymbol{a} = x + 1$, always makes a small positive contribution to the impact; but

(iii) adding one more contribution, so that $\sum \boldsymbol{a} = x + 1$, never causes a change in experienced harm compared to x.

Contribution is Wrong. Any action that contributes to the impact is wrong.

Direct Act Consequences, **Experienced Harm**, **No Stepwise Harm** and **Contribution is Wrong** are inconsistent, because (from **No Stepwise Harm**) there exist actions that do not cause experienced harm while holding all other actions fixed, even though they contribute to the impact. From **Direct Act Consequences** we know that we should only consider the immediate consequences of an action to determine whether it is wrong, holding all other actions fixed. Therefore (from **Experienced Harm**) these actions do not have the wrong-making property of causing (expected) harm and are not wrong. This contradicts **Contribution is Wrong**.

How plausible are these propositions? Take **Direct Act Consequences** first. Act consequentialists are committed to the claim that the moral properties of an action depend on its consequences only. In fact, all reasonable ethical theories would take the consequences of actions, perhaps among other considerations, into account. The more controversial aspect of **Direct Act Consequences** is the claim that only the consequences of *this particular action, holding all other relevant actions fixed*, should be considered. The plausibility of this more stringent condition stems from considerations of control. To assess what is under the control of the acting person, we need to know what exact difference this person's action can make to the world. We work this out by comparing the consequences of performance and non-performance, while holding everything else fixed.[4] **Direct Act Consequences** ensures that the rightness or wrongness of one's action depends only on the consequences one has control over.

Experienced Harm is a commitment to a specific conception of harm. According to **Experienced Harm**, one property that is required to make an action wrong is the expectation of experienced harm, or, more precisely, the *phenomenal* component of harm. The most obvious candidates here are pain and discomfort. Unsurprisingly, classical hedonist act-utilitarians will have no issue with **Experienced Harm**. However, even if most philosophers are probably unwilling to subscribe to **Experienced Harm** across the board, **Experienced Harm** may at least be plausibly applicable in those situations where no other bad effects can be identified except for the harm experienced by the victim.

Let us turn to **No Stepwise Harm**. If we assume **No Stepwise Harm**, each single additional contribution leads to an increase in the impact m but will not lead to an increase in experienced harm. How can this be? The explanation is that the experience of harm is *minimal change insensitive* (short: *insensitive*), such that adding or subtracting just one contribution never leads to a change in the perception of the current level of harm (here I am inspired by the formal treatment of the related concept of vagueness in van Rooij, 2010).

To express this a little more technically, assume that we order the harms according to the relations $\succ_H$ ("is experienced as more harmful than") and an

[4] How precisely the counterfactual is to be cashed out is a surprisingly difficult question, but here a rough sketch suffices. See Norcross (2005).

indifference relation $\sim_H$ ("is experienced as equally harmful as"). According to **No Stepwise Harm**, clause (iii), there exist decision situations such that any pair of harms that differ by just one additional or omitted contribution is indistinguishable. This results in a series of indistinguishable harm observation pairs:

$$h(m_{c=0}) \sim_H h(m_{c=1}) \wedge h(m_{c=1}) \sim_H h(m_{c=2}) \wedge \ldots \wedge h(m_{c=n-1}) \sim_H h(m_{c=n}). \quad (1)$$

It is now clear that the relation $\sim_H$ cannot be an equivalence relation. If it was, it would follow immediately that $h(m_{c=0}) \sim_H h(m_{c=n})$, which would contradict clause (i) of **No Stepwise Harm**. Of course, some have taken this as a *reductio*, claiming that a premise like **No Stepwise Harm** must be rejected (e.g., Graff, 2001). I will argue, however, that the insensitivity involved here makes it plausible to accept **No Stepwise Harm**.

For **No Stepwise Harm** to be consistent, we need to explain how the claim in clause (iii) can be true without directly contradicting (i). The key to this is the assumption that the relation $\sim_H$ can be *non-transitive*.[5] Therefore, all the conjuncts of (1) can be true, but, due to non-transitivity, this does not imply that $h(m_{c=0}) \sim_H h(m_{c=n})$ and avoids the contradiction with (i). In other words: it is possible that adding or taking away the contribution of one individual never increases or lowers the experienced harm, but putting all (or more than one) actions together does.[6]

The vagueness[7] literature contains several proposals as to how relations like "is indistinguishable from", "feels just as", and so on, can be non-transitive. Here I state just one such proposal. According to one type of "contextualist" account, the relation of indistinguishability depends on context, and the context changes in a subtle way while moving through the series. For instance, if the relation "is experienced as equally harmful as" is a slightly different relation for some levels of contribution,

[5] More technically, van Rooij (2010) and others before him account for this with semi-orders. Note that (1), together with the assumption of non-transitivity, does *not* give rise to the sorites paradox because (1) and non-transitivity does not imply the inductive premise needed for a sorites problem. I will revisit this point below.

[6] Mike Otsuka (1991, p. 138), for example, explicitly endorses the view that such situations exist.

[7] While the literature on vagueness inspires my approach, I prefer the term "insensitivity" to mark the fact that vagueness pertains to semantics, while insensitivity pertains to perception. I am grateful to Christian List for urging me

then some of the relations in (1) are slightly different, and if some of the relations are different, transitivity cannot be applied across them all (e.g., Raffman, 1996). After all, transitivity is a property of one and the same relation, it has no bearing on a series of subtly different relations. If we can accept such an interpretation of clause (iii), then **No Stepwise Harm** can plausibly be true. Further evidence that this is a plausible assumption comes from the literature on phenomenal sorites problems. Many philosophers[8] engaged in this debate endorse the view that it is possible to experience insensitivity in terms of color perception. If two color patches are *very* similar (even though we know from the way we produced them that they do not have exactly the same hue), then subjects cannot distinguish between them in pairwise comparison. A very similar phenomenal experience is plausible with regard to pain perception.

Last, we look at **Contribution is Wrong**. This expresses a fairly common intuition about doing harm. If you play a contributing part in a causal process that harms people by increasing the impact, then this contributing action is wrong. This, according to **Contribution is Wrong**, is independent from whether your action alone makes a perceptible difference, it suffices that you positively contribute to the impact.

To avoid the contradiction between **Direct Act Consequences**, **Experienced Harm**, **No Stepwise Harm** and **Contribution is Wrong**, we need to weaken or give up at least one of these four propositions. Using this framework, we can analyse the recent debate on imperceptible effects and "many hands problems". Different participants have proposed to relax different propositions, but very few have carefully considered the underlying background assumptions and possible trade-offs. The next four sections account for different attempts to relax one of the four propositions to avoid the contradiction. I sketch my preferred solution in section 8.

4 Relaxing Direct Act Consequences

In *Reasons and Persons*, Parfit defines a view very similar to **Direct Act Consequences**, which he calls the Second Mistake:

---

to make this distinction.

[8] Graff (2001) reviews the main players, but comes to endorse transitivity.

> "(The Second Mistake) If some act is right or wrong *because of its effects*, the only relevant effects are the effects of this particular act." (Parfit, 1984, p. 70)

Parfit argues that this view is mistaken because it leads to implausible implications in overdetermination cases. To demonstrate, Parfit discusses the famous "Two Assassins". If two snipers shoot at me at exactly the same time and the two bullets pierce my heart such that each of them would have been sufficient to kill me immediately, then each of the two assassins alone is not causally necessary for killing me.[9]

Parfit's preferred solution for problems of this sort is to suggest that even if individual acts do not produce harm, it is possible that individuals *together* can create harms:

> "(C7). Even if an act harms no one, this act may be wrong because it is one of a *set* of acts that *together* harm other people." (Parfit, 1984, 70)[10]

This principle states that acts can be wrong because they are part of a set of acts that causes harm, in contradiction to **Direct Act Consequences**. Even though one single act may be 'harm-less' in the sense of **No Stepwise Harm**, the set of acts is not. For instance, even if Sinnott-Armstrong's pleasure driver is not causing any harm individually, his driving is part of a set of actions that do cause harm, namely all the acts leading to greenhouse gas emissions. Therefore, relaxing **Direct Act Consequences** by suggesting a principle along the lines of C7 has some appeal.

One difficulty with notions like "harming together" is to deliver a plausible explanation why the *individual* acts are impermissible if they cause harm *together* (Tannsjö, 1989, p. 223). In addition, note that according to C7 an individual act is not *necessarily* wrong just because it is part of a harming set of acts. Parfit aims to

[9] Not everyone agrees with Parfit that the Second Mistake really is a mistake. Frank Jackson (1997) maintains that act-consequentialists should hold their nerve: In cases of overdetermination, they should maintain that the single acts are not wrong because they are not causally necessary to bring about the harm. If one wants to defend Jackson's line, it is important to be precise about the setup. In Jackson's discussion of the Two Assassins, it is *certain* that both of the Two Assassins will shoot, and they shoot independently from each other. In such a case, the action of one assassin alone truly does not cause any harm.

qualify C7, since in his view, what is required of individuals depends on what the other individuals do. People are not required to withhold actions if there is no harm in the aggregate, or if the harm will happen anyhow. This is why, in *Reasons and Persons*, Parfit conditions the obligation on enough but not too many others acting in a similar way and introduces a common knowledge assumption (see C10 on p. 77). In addition, as both Gruzalski (1986) and Otsuka (1991) point out, Parfit's "together" solution crucially depends on two inconsistent claims: that one can determine a most beneficial (and equivalently: least harming) set of actions, while at the same time assuming that adding or withdrawing one contribution makes no difference. These problems remain a challenge for Parfit's view.

Overall, Parfit's proposal for relaxing **Direct Act Consequences**, while clearly intuitively attractive, has faced objections that have not been fully addressed. Below I will suggest a different way to relax **Direct Act Consequences** that, I hope to show, holds more promise.

5 Relaxing the Experienced Harm Assumption

Rejections of **Experienced Harm** are frequent in the literature. Glover suggests that actions without experienced harm as a consequence are still wrong because they must be understood as producing *fractions* of perceptible harm:

> "[The Principle of Divisibility] says that, in cases where harm is a matter of degree, subthreshold actions are wrong to the extent that they cause harm, and where a hundred acts like mine are necessary to cause a detectable difference I have caused 1/100 of that detectable harm." (Glover and Scott-Taggart, 1975, p. 174)

Unfortunately, Glover's argument for the Principle of Divisibility is rather weak: he simply points to the unpalatable implications of rejecting that principle.

More promising are arguments that give us an explanation why actions that do not cause experienced harm can be wrong nevertheless. Several authors have observed that even if the addition of one more actions may not change the perceived harm, at some point the *report* of the overall level of pain must change. In Parfit's torturer

[10] I quote (C7) rather than (C10), as Parfit has retracted on (C10) in his reaction to an argument presented by Gruzalski. See Parfit (1986) and Gruzalski (1986).

case, every two situations that differ by only one step in the increase of voltage feel exactly the same when compared against each other, but at the same time it must be true that the subjects' experience of pain changes when the voltage gets notched up gradually. If absolutely *no* property changed while the voltage increases, it could not be true that there is no pain at voltage 0, but tremendous pain at voltage 1000V. The question is which property one should consider to avoid the problem. Kagan considers the subject's pain report:

> "At some point the answer to the question "are you in pain?" must differ from the answer given immediately before—otherwise the victim would still be answering "no" at state 1,000 (just as they answered "no" at state 0), something we know to be false." (Kagan, 2011, p. 132)

Kagan thus proposes to take the *pain report* as the relevant property. It is undoubtedly true that when the number of contributions c increases, at some number of contributions x the impact has increased enough for the experienced harm to be greater than the (lack of) harm at the beginning, that is to say $h(m_0) \prec_H h(m_x)$. From this Kagan infers that "at least one state must feel different from the one that came before" (Kagan, 2011, p. 132). This last claim, however, is mistaken, and it contradicts what Kagan says earlier about such cases: "each individual act makes no perceptible difference to anyone's pain" (p. 115).

Kagan begins with a careful definition of the morally relevant measure, the perception of pain, only to gradually conflate it with other measures. The fact that *reported* pain must eventually change between two steps in the increase of voltage simply does not imply that the experienced pain is different between these two steps. The subject's report can consistently look like this:

…

Step x : Does the pain feel just like at step $x-1$? "Yes"

Are you in pain? "No"

Step $x+1$**:** Does the pain feel just like at step x? "Yes"

Are you in pain? "Yes"

At step $x+1$ the subject may be thinking: "This feels just like the last step. But it really does feel painful now, so I say yes to the second question." Despite Kagan's

best efforts to convince us otherwise, the subject does not make a mistake in reporting their perception (Nefsky, 2011). What would be mistaken is the belief that *nothing* has changed – but that does not entail that the subject must therefore be able to distinguish the pain perception of step x and $x + 1$ in direct comparison.

Similar claims can be defended in analogous settings. Arguably, one can gradually change the temperature such that a subject can never tell the difference between any two notches, one can gradually fade from one color to another such that each two adjacent shades are indistinguishable, and so on. This does not prevent the subjects from realizing along the way that certain variables have changed. They begin to notice pain, they notice that the water that was cold at the beginning is now lukewarm, or that the green is now orange. What they cannot do is distinguish between adjacent steps. Consequently, they also cannot pinpoint the precise step where the change has taken place – even though they do notice the change.

What Kagan effectively proposes is a revision of **Experienced Harm**. Instead of considering experienced harm, he suggests considering *harm reports*. And since contributions are typically made under conditions of imperfect information, any contribution has a small probability to change the harm reported, such that the *expected reported harm* is positive for all contributions.[11]

Another way to revise **Experienced Harm** is to appeal to properties that are accessible by theoretical reasoning (Hansson, 1999; Shrader-Frechette, 1987). Provided that the subjects are well-informed about the setup, they know that even though they cannot feel a difference in pain when the contribution level changes slightly, they still know that the impact on them has changed. In case of the harmless torturers one could think about the subjects sitting in front of a voltmeter with very high measurement resolution, displaying the current voltage. One contribution is added. The subjects do not feel any difference, but they do see the increased voltage on the voltmeter. From this they infer that it has become more likely that their pain perception will change once further contributions are added.

Extending our limited cognitive abilities by theoretical reasoning or by measurement instruments is common in science and in everyday life. Suppose I cut a piece of butter into two equal halfs. They feel equally heavy to me. However,

[11] A related, probabilistic version along the same lines was given earlier by Voorhoeve and Binmore (2006) and by Arntzenius and McCarthy (1997).

theoretical reasoning tells me that with a probability infinitely close to 1, one piece must be a tiny bit heavier than the other, and a sufficiently precise scale will tell me which one this is. It would be crazy to maintain that we should end the inquiry by insisting that the two lumps of butter are equally heavy because they feel equally heavy. And perhaps we should also use suitable reasoning and measurement tools to get a better empirical access to the causal process that leads to the harm the subjects are experiencing in our problem. That means that we should stop taking phenomenally experienced harm as the relevant wrong-making property; instead we should consider variables such as "contribution to the impact leading to harm".

I am broadly sympathetic to the proposal to replace **Experienced Harm** with a more theoretically and empirically informed measurement of harm. It avoids the inconsistency by correcting the limitations of our insensitivity in perception in a theoretically compelling way. However, as attractive as this solution is, it also comes with a challenge: a moral theory that departs from **Experienced Harm** must explain what makes an action wrong if no one experiences any consequences caused by this action. The challenge is complicated by the fact that not all contributions to an impact that causes harm are obviously wrong. For instance, suppose the torturers push their buttons sequentially. Let it be a known fact that 999 of the harmless torturers have already pushed their button. Is it wrong for the final torturer to also push her button? The action of the last torturer will not lead to any difference in pain, but it does add to the impact (in voltage). For what it's worth, my intuition is that the action of the last torturer is irrelevant and therefore not wrong. However, that sort of intuition does not sit well with with more theoretically informed measurements of harm, as just sketched. My preferred solution, described in section 8, is better suited to deal with such cases.

6 Relaxing the No Stepwise Harm Assumption

The only sub-clause of **No Stepwise Harm** that can plausibly be relaxed is (iii): that any single additional action does not cause any experienced harm, given all other actions.

A first revision would move from perceived harm to *perceivable* harm.[12] Proponents of this revision can concede that the subjects do not *actually* experience any difference in pain when one contribution is added or removed. They do maintain, however, that the subjects are mistaken in claiming that the pain is the same, and that the subjects would be able to notice the difference under ideal conditions of perception. Therefore, there is a perceivable, though unperceived, difference in harm whenever the level of contribution changes. I am unsure how to understand this argument. It may be that the difference becomes perceivable by using theoretical knowledge or tools of measurement as discussed in the previous section. If that is the case, the proposal looks like a revision of **Experienced Harm**, as discussed above. Or it may be that the subjects perceive the difference without awareness, but are able to become aware of the differences (Mills, 2002, p. 392). That latter claim is an empirical speculation, and there is no reason to believe that it would hold for extremely small differences in impact.

A second revision of **No Stepwise Harm** consists in contextualising the comparisons between settings. Proponents of this solution are prepared to admit that changes of just one contribution do not cause additional experienced harm. However, for them that is beside the point. The harm increase from x contributions to $x+1$ contributions will be noticed, it is claimed, once we compare these settings not only pairwise against each other, but with other levels of contributions. For example, Voorhoeve and Binmore suggest that

> "Two adjacent notches might be indistinguishable in this way because [the subject's] pain experience at a particular notch might (because of some neurophysiological process that we need not understand) depend on the current she was exposed to before. Thus, it might be the case that if the previous current is very different, [the subject] experiences the current at notch n in one way, but if it is similar, (i.e. the difference between them is smaller than the just-noticeable difference) she experiences it in another way." (Voorhoeve and Binmore, 2006, p. 105)

[12] An argument along this line was suggested but not necessarily endorsed by Gunnar Björnsson (personal communication).

No matter how exactly the described phenomenon would be caused, Voorhoeve and Binmore want to suggest that the reported indistinguishability in pairwise comparisons is a mistake, and that this mistake could be avoided if, first, settings with quite different voltages are compared, and second, many such settings are tried, so that one can establish the frequencies of different pain reports. In effect, Voorhoeve and Binmore propose to revise both **No Stepwise Harm** and **Experienced Harm** to appeal to reflectively corrected harm judgements.

The question remains, however, why individuals ought to be compelled to make this rationality adjustment. To be clear, Voorhoeve and Binmore's proposal is not meant to address many hands problems but settings that threaten *individual* rationality, especially Quinn's famous "self torturer" (Quinn, 1990). Getting what is worst for yourself through your own choices clearly is a failure of individual rationality. It makes sense to try to avoid this failure by rational reflection. This reflection will reveal that you ought not to prefer each one-step increase of voltage and you ought to correct your preference structure accordingly. In the inter-personal case of many hands, however, this line of reasoning does not apply so easily because the harm is neither self-inflicted nor caused by one and the same agent. Perhaps one could claim that despite forming judgements of pain perception with a non-transitive "is experienced as equally harmful as" relation, the victim ought to form *transitive preferences over the outcomes that can arise*. And if that is so, one could say that the torturers ought to orientate their actions according to these transitive preferences. However, this line of argument moves towards a quite different form of consequentialism, a "preference consequentialism" that requires a complete revision of **Experienced Harm** as well.

7 Relaxing the Wrongness of Contribution Assumption

Contribution is Wrong links contributions to wrongness, even if the individual action does not lead to experienced harm.[13] Relaxing this principle does not, on first sight, appear to be a promising way out.

[13] However, actions that do not cause any change in impact are not necessarily wrong according to **Contribution is Wrong**.

One caveat applies: there may be settings in which the contribution to a causal process leading to harm is such that the rest of the causal process is entirely fixed. For example, assume that Bob sits on a dyke in a major flood area. Huge amounts of waters are flowing, foreseeably leading to a humanitarian catastrophe further downstream. The amount of water flowing is due entirely to natural processes, and we can assume that it is a fixed amount. Also assume that tipping point effects can be ruled out. In this situation Bob pours one pint of water into the floods. Bob's contribution is so small that it is not leading to any perceptible harm. Is it wrong for Bob to pour his pint?

According to **Contribution is Wrong**, Bob minimally contributes to the impact, i.e. the volume of water that causes harm downstream, and his action is therefore wrong. However, in such cases my intuition, at least, pulls in the other direction. For reasons that I will explore in greater detail in the next section, it makes a difference whether the causal process one deals with is fixed or the result of agential choices.

8 The Sketch of a Solution Proposal

At the heart of the hardest many hands problems lies the insensitivity of harm perception. Taking this problem seriously means acknowledging that minimal increases in impact may not be perceptible. And if the effects of single actions are not perceptible, it *seems* to follow that the aggregate effects of many such actions are not perceptible either. I say "*seems*" because I think that this last step is mistaken. Put very roughly, it is mistaken because the imperceptibility of single actions does not imply that the aggregate effects of many actions are equally imperceptible. The error occurs because many have thought that the relation $\sim_H$ ("is experienced as equally harmful as") must be transitive. But this need not be the case. In this section I lay out the steps of the argument just sketched more carefully and discuss its normative implications.

Readers familiar with the sorites paradox literature will have noted that **No Stepwise Harm** does *not* set up a sorites paradox. To get the paradox, a stronger inductive premise is needed:

Base Premise. When 0 torturers push the button, the victim is not in pain.

Inductive Premise. If the victim is not in pain if x torturers push the button, then the victim is not in pain if $x + 1$ torturers push the button.

Conclusion. The victim is not in pain if all torturers push the button.

The paradox arises because we arrive at the absurd conclusion that the harmless torturers are indeed – harmless. A plausible diagnosis of the argument is that the Inductive Premise is false. However, clause (iii) in **No Stepwise Harm** is subtly weaker. The formalization from above makes this quite clear. **No Stepwise Harm** only postulates that the difference in each pairwise comparison between the pain at step x and at step $x + 1$ is imperceptible. This results in a series of pairwise relations:

$$h(m_{c=0}) \sim_H h(m_{c=1}) \wedge h(m_{c=1}) \sim_H h(m_{c=2}) \wedge \ldots \wedge h(m_{c=n-1}) \sim_H h(m_{c=n}). \qquad (1)$$

The inductive premise, by contrast, implies this:

$$h(m_{c=0}) \sim_H h(m_{c=1}) \sim_H h(m_{c=2}) \sim_H \ldots \sim_H h(m_{c=n-1}) \sim_H h(m_{c=n}). \qquad (2)$$

(1) entails (2) only if the relation $\sim_H$ is transitive. But it is the transitivity of $\sim_H$ that I deny.

Transitivity is often assumed without explicit argument because we tend to think about similarity relations in terms of equivalence relations or as the symmetrical part of a weak order. Moreover, transitivity is a fundamental rationality constraint for preferences. In the context of preferences it should not be abandoned lightly. In the context of perception, however, it is far from clear whether transitivity in similarity judgements obtains empirically. Also, unlike in the preference case, it is not a fundamental rationality constraint. Thus, assuming non-transitivity for similarity of harm perceptions is much less objectionable.

Two implications follow immediately. First, we can now account for the plausible claim that individuals cannot perceive a difference in harm if the difference in impact is very small. Second, this does not lead to a sorites paradox because the non-transitive pairwise rankings in (1) are weaker than the inductive premise required to set up the paradox. In particular, from the fact that $h(m_{c=x}) \sim_H h(m_{c=x+1})$ for any $x \in (0, n-1)$ it does *not* follow that $h(mc_{c=x}) \sim_H h(m_{c=x+2})$ for any

$x \in (0, n-2)$. The fact that small steps are indistinguishable does not imply that larger steps are indistinguishable.

The last point has an important upshot: if larger steps are distinguishable in their harm, then aggregate actions can lead to perceptible harm even if they consist of many small actions that are individually not harmful. This opens up a potentially attractive solution to avoid the contradiction between **Direct Act Consequences**, **Experienced Harm**, **No Stepwise Harm** and **Contribution is Wrong**. More specifically, what is required is a revision of **Direct Act Consequences** such that individuals take into account how their action can be effective when it occurs with other actions, rather than keeping all other actions fixed. Here I will simply sketch the rough shape of the revision required, a detailed solution proposal and a thorough defence will have to wait for another time.

It is clear that each individual on their own, given the contributions of all other individuals, cannot change the harm perceived by the victim. At the same time, it is the case that several individuals *together* can change the victim's pain (this, of course, is also the motivation behind Parfit's solution). Therefore, it is useful to consider *minimal perceptible subsets* of the set of all actions. Minimal perceptible subsets are those sets that contain just enough actions such that together these actions avoid minimal change insensitivity – they can jointly make a perceptible difference. More specifically, if all actions in a minimal perceptible subset change from "not contribute" to "contribute" (or vice versa), the victim notices a difference.[14]

Consider individual *i*'s action a_i. This action is an element in some of the minimal perceptible subsets. Each set can make a difference to harm if all actions in the set change from "not contribute" to "contribute" (and vice versa). Action a_i therefore contributes to expected harm together with others in a minimal perceptible subset if (i) a_i is "contribute" and (ii) there is a positive ex ante probability that all others actions in the set (except a_i) are "contribute" and a positive probability that they are all "not contribute". Condition (i) checks whether the individual contributes; condition (ii) whether there is a positive probability that the minimal perceptible subset can make a difference.

[14] My solution is inspired by Braham and van Hees's (2009, 2012) treatment of moral and causal responsibility, though my sketch solution differs from theirs and my framework is much less general.

Crucially, whether a_i is in the position to increase expected harm depends on the joint probabilities of the other actions. If a_i cannot make a difference to the consequence this subset of actions has because, for example, other actions are fixed, then a_i does not increase expected harm by contributing with regard to that set. And if a_i cannot contribute to a difference in any of the minimal perceptible subsets it is an element of, then it does not increase expected harm at all.

A solution along the lines sketched here will have to revise **Direct Act Consequences** by incorporating a non-standard conception of consequences. This non-standard conception will have to say that contributing in suitable minimal perceptible subsets has the consequence of increasing expected experienced harm. This, in turn, will allow us to say that actions that are individually imperceptible are still wrong because they are expected to bring about harm together with other actions.

For example, the action of pushing a button in the harmless torturers setting leads to increased expected experienced harm for the victim. Even though the action does not *by itself* increase experienced harm, the action is expected to do so in the following sense: the action (to contribute to the voltage) will be part of minimal perceptible subsets of contributions such that the actions of these subsets together can be felt by the victim. And since there is a positive probability that all the actions in the subsets are "not contribute" or "contribute", there is a positive probability that these subsets can make a difference. For instance, suppose that the victim always experiences additional pain if two more torturers push the buttons, but never if just one additional torturer pushes the button. This means that the button pushing of one torturer is part of subsets with two elements that can be felt. And since there is a positive probability for a subset to make a difference in harm, contributing within such a subset leads to expected harm when adopting the non-standard notion of consequences suggested in this section.

The solution sketched offers a suitably differentiated account of cases where the choices of other agents are either truly agential free choices, and cases where the other contributions are fixed (and therefore not subject to choice). Suppose a single torturer shows up for work (cf. Parfit, 1984, p. 81) and, as introduced above, 999 buttons are already pushed so that the victim suffers severe pain. If pushing or not pushing the button is the only available action, is it permissible to push one more button? Parfit is inclined to say "no", as he thinks that the "second mistake" is indeed

that: a mistake. But according to my proposal, the single push of the button has a zero expectation to contribute to harm, and is therefore permissible. The reason is that while such an action is a member of minimal perceptible subsets, the probability that any of these subsets will jointly make a perceptible difference is zero, as all the other actions are fixed. Similar implications follow for other determined processes. If one pours a pint of water into a devastating flood and the amount of floodwater is already fully determined, this action is permissible. But it is not permissible if the amount of floodwater is not fully determined and one's pint could with some non-zero probability make a difference in combination with other stochastic processes (and this is the more realistic setting in the real world). These examples show that the solution I sketch here will also have to revise **Contribution is Wrong**: contributions are only wrong if they cause an increase in expected experienced harm, if they cannot make a difference to expected harm they are permissible.

Unlike Kagan, I do not rely on any actual thresholds being crossed by any specific contributing individual. In particular, I do not have to claim that, at some point, the action of one individual must trigger harm. And, again unlike Kagan (p. 130), I do not have to deny the existence of genuine imperceptible effects cases. In fact, my proposal shows that Kagan's claim "that there could not possibly be cases of imperceptible difference" (p. 130) is false. There can be cases of imperceptible difference if the relevant relation is non-transitive because of insensitivity. My solution can therefore preserve the empirically and conceptually plausible assumption that there exist changes in impact so small that they never register as additional harm. Towards the end of the paper, Kagan admits that the differences from notch to notch may not be "directly" (p. 137) perceivable. If he weakens his view in that way, however, he concedes that one single action can be imperceptible, undermining his central conceptual claim. In any case, in his final discussion Kagan vacillates between denying **No Stepwise Harm** and denying **Experienced Harm**.

9 Conclusion

The hardest many hands problems are those where each individual action does not register in terms of perceived harm, but collectively the actions are very harmful. The solution I propose requires individuals who find themselves in such a situation to consider what the consequences of their action might be in combination with other

actions. I do not simply claim that contributing individuals do wrong because they are part of a harming group. The contributing individuals do wrong because they ignore the risk that their action, even though it cannot be perceived while holding all other actions fixed, may well be perceived with others if we do not hold everything else fixed. The individual action can become perceivable in a set with other actions and therefore contributes to expected harm. I claim that performing actions that might become phenomenally effective together with others is wrong because the individual's action increases the chance that such a harm comes about. To avoid doing wrong, individuals must be forward-looking and think about how their contributions to the impact might lead to experienced harm.

References

Andreou, Chrisoula. 2006. "Environmental Damage and the Puzzle of the Self-Torturer." *Philosophy & Public Affairs* 34(1):95–108.

Arntzenius, Frank and David McCarthy. 1997. "Self Torture and Group Beneficence." *Erkenntnis* 47(1):129–144.

Braham, Matthew and Martin van Hees. 2009. "Degrees of Causation." *Erkenntnis* 71(3):323–344.

Braham, Matthew and Martin van Hees. 2012. "An Anatomy of Moral Responsibility." *Mind* 121(483):606–634.

Broome, John. 2012. *Climate matters: ethics in a warming world*. New York: W.W. Norton.

Glover, Jonathan and M. J. Scott-Taggart. 1975. "It makes no difference whether or not I do it." *Proceedings of the Aristotelian Society, Supplementary Volumes* 49:171–209.

Graff, Delia. 2001. "Phenomenal Continua and the Sorites." *Mind* 110(440):905–936.

Gruzalski, Bart. 1986. "Parfit's Impact on Utilitarianism." *Ethics* 96(4):760–783.

Hansson, Sven Ove. 1999. "The Moral Significance of Indetectable Effects." *Risk: Health, Safety & Environment* 10:101–108.

Jackson, Frank. 1997. Which Effects? In *Reading Parfit*, ed. Jonathan Dancy. Oxford: Blackwell pp. 42–53.

Kagan, Shelly. 2011. “Do I Make a Difference? ” *Philosophy & Public Affairs* 39(2):105–141.

Lichtenberg, Judith. 2010. “Negative Duties, Positive Duties, and the "New Harms".” *Ethics* 120:557–578.

Mills, Eugene. 2002. “Fallibility and the phenomenal sorites.” *Noûs* 36(3):384–407.

Nefsky, Julia. 2011. “Consequentialism and the Problem of Collective Harm: A Reply to Kagan.” *Philosophy & Public Affairs* 39(4):364–395.

Norcross, Alastair. 2005. “Harming In Context.” *Philosophical Studies* 123(1-2):149–173.

Otsuka, Michael. 1991. “The Paradox of Group Beneficence.” *Philosophy & Public Affairs* 20(2):132–149.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Clarendon.

Parfit, Derek. 1986. “Comments.” *Ethics* 96(4):832–872.

Quinn, Warren S. 1990. “The puzzle of the self-torturer.” *Philosophical Studies* 59(1):79–90.

Raffman, Diana 1996. “Vagueness and context-relativity.” *Philosophical Studies* 81:175–192.

Shrader-Frechette, Kristin. 1987. “Parfit and Mistakes in Moral Mathematics.” *Ethics* 98(1):50–60.

Sinnott-Armstrong, Walter. 2005. It’s Not My Fault: Global Warming and Individual Obligations. In *Perspectives on Climate Change: Science, Economics, Politics, Ethics*, ed. Walter Sinnott-Armstrong & Richard B Howarth. Amsterdam: Elsevier pp. 285–307.

Tannsjö, Torbjörn. 1989. “The Morality of Collective Actions.” *The Philosophical Quarterly* 39(155):221–228.

Thompson, Dennis F. 1980. “Moral Responsibility of Public Officials: The Problem of Many Hands.” *The American Political Science Review* 74(4):905–916.

van Rooij, R. 2010. “Vagueness, tolerance, and non-transitive entailment.” Manuscript. Available at http://web.logic.at/lomorevi/vaguebook/rooij.pdf.

Voorhoeve, Alex and Ken Binmore. 2006. “Transitivity, the Sorites paradox, and similarity-based decision-making.” *Erkenntnis* 64(1):101–114.