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.