The Unilateralist's Curse: The Case for a Principle of Conformity

Nick Bostrom
Anders Sandberg
Tom Douglas

[(2013) Working paper, Future of Humanity Institute, Oxford University]
www.nickbostrom.com

ABSTRACT. In some situations a number of agents each have the ability to undertake an initiative that would have significant effects on the others. Suppose that each of these agents is purely motivated by an altruistic concern for the common good. We show that if each agent acts on her own personal judgment as to whether the initiative should be undertaken, then the initiative will move forward more often than is optimal. We suggest that this phenomenon, which we call *the unilateralist's curse*, arises in many contexts, including some that are important for public policy. To lift the curse, we propose a *principle of conformity*, which would discourage unilateralist action. We consider three different models for how this principle could be implemented, and respond to some objections that could be raised against it.

1. Introduction

Consider the following hypothetical scenarios:

> 1. A group of scientists working on the development of an HIV vaccine have accidentally created an airborne transmissible variant of HIV. They must decide whether to publish their discovery, knowing that it might be used to create a devastating biological weapon, but also that it could help those who hope to develop defenses against such weapons. Most members of the group think publication is too risky, but one disagrees. He mentions the discovery at a conference, and soon the details are widely known.
>
> 2. A sports team is planning a surprise birthday party for its coach. One of the players decides that it would be more fun to tell the coach in advance about the planned event. Although the other players think it would be better to keep it a surprise, the unilateralist lets word slip about the preparations underway.
>
> 3. Geoengineering techniques have developed to the point that it is possible for any of the world's twenty most technologically advanced nations to substantially reduce the earth's average temperature by emitting sulfate aerosols. Each of these nations separately considers whether to release such aerosols. Nineteen decide against, but one nation estimates that the benefits of lowering temperature would exceed the costs. It presses ahead with its sulfate aerosol program and the global average temperature drops by almost 1 degree.

It is plausible that, in each of these cases, each of a number of agents is in a position to undertake an initiative, *X*. Each agent decides whether or not to undertake *X* on the basis of her own independent judgment of the value of *X*, where the value of *X* is assumed to be independent of *who* undertakes *X*, and is supposed to be determined by the contribution of *X* to the common good.[1] Each agent's judgment is subject to error—some agents might overestimate the value of *X*, others might underestimate it. If the true value of *X* is negative, then the larger the number of agents, the greater the chances that at least one agent will overestimate *X* sufficiently to make the value of *X* seem positive. Thus, if agents act unilaterally, the initiative is too likely to be undertaken, and if such scenarios repeat, an excessively large number of initiatives are likely to be undertaken. We shall call this phenomenon the *unilateralist's curse*.

Though we have chosen to introduce the unilateralist's curse with hypothetical examples, it is not merely a hypothetical problem. There are numerous historical examples, ranging from the mundane to the high-tech. Here is one:

> Until the late 1970s, the mechanism of the hydrogen bomb was one of the world's best kept scientific secrets: it is thought that only four governments were in possession of it, each having decided not to divulge it. But staff at the Progressive magazine believed that nuclear secrecy was fuelling the Cold War by enabling nuclear policy to be determined by a security elite without proper public scrutiny. They pieced together the mechanism of the bomb and published it in their magazine, arguing that the cost, in the form of aiding countries such as India, Pakistan and South Africa in acquiring hydrogen bombs, was outweighed by the benefits of undermining nuclear secrecy.[2]

It is perhaps too soon to say whether this was the wrong decision. But in other cases, it is clearer that unilateral action led to a suboptimal outcome:

> In the mid-nineteenth century there were virtually no wild rabbits in Australia, though many were in a position to introduce them. In 1859, Thomas Austin, a wealthy grazier, took it upon himself to do so. He had a dozen or two European rabbits imported from England and is reported to have said that "The introduction of a few rabbits could do little harm and might provide a touch of home, in addition to a spot of hunting."[3] However, the rabbit population grew dramatically, and rabbits quickly became Australia's most reviled pests, destroying large swathes of agricultural land.[4]

2. The unilateralist's curse: a model

The unilateralist's curse is closely related to a problem in auction theory known as the winner's curse. The winner's curse is the phenomenon that the winning bid in an auction has a high likelihood of being higher than the actual value of the good sold.[5] Each bidder makes an independent estimate and the bidder with the highest estimate outbids the others. But if the average estimate is likely to be an accurate estimate of the value, then the winner overpays. The larger the number of bidders, the more likely it is that at least one of them has overestimated the value.

The unilateralist's curse and the winner's curse have the same basic structure. The difference between them lies in the goals of the agents and the nature of the

decision. In the winner's curse, each agent aims to make a purchase if and only if doing so will be valuable *for her.* In the unilateralist's curse, the decision-maker chooses whether to undertake an initiative with an eye to the common good, that is, seeking to undertake the initiative if and only if the initiative contributes positively to the common good.

The unilateralist's curse can be illustrated using a simple mathematical model. Assume N agents, each considering whether to undertake an initiative. Each agent wishes to proceed if and only if the value of the initiative is positive, but the agents do not know the true value V^* of the initiative (which may be negative or positive). Instead each agent forms an estimate that is the sum of V^* and a random independent error d drawn from a distribution with cumulative distribution function $F(d)$. This means that the probability p that any given agent will estimate the value of the initiative to be positive when it is in fact negative ($V^*<0$) is $p = 1 - F(-V^*)$.[6] The probability P that at least one of the agents will incorrectly estimate the value to be positive is $P = 1 - (1 - p)^N = 1 - F(-V^*)^N$.

For the case with 5 agents and d as a random error drawn from a normal distribution with standard deviation 1 and mean zero, the probability that any initiative will be undertaken (regardless of whether it is a good idea or not) is high even when the true value is quite negative and the probability rises steeply as the true value of the initiative approaches zero from below.

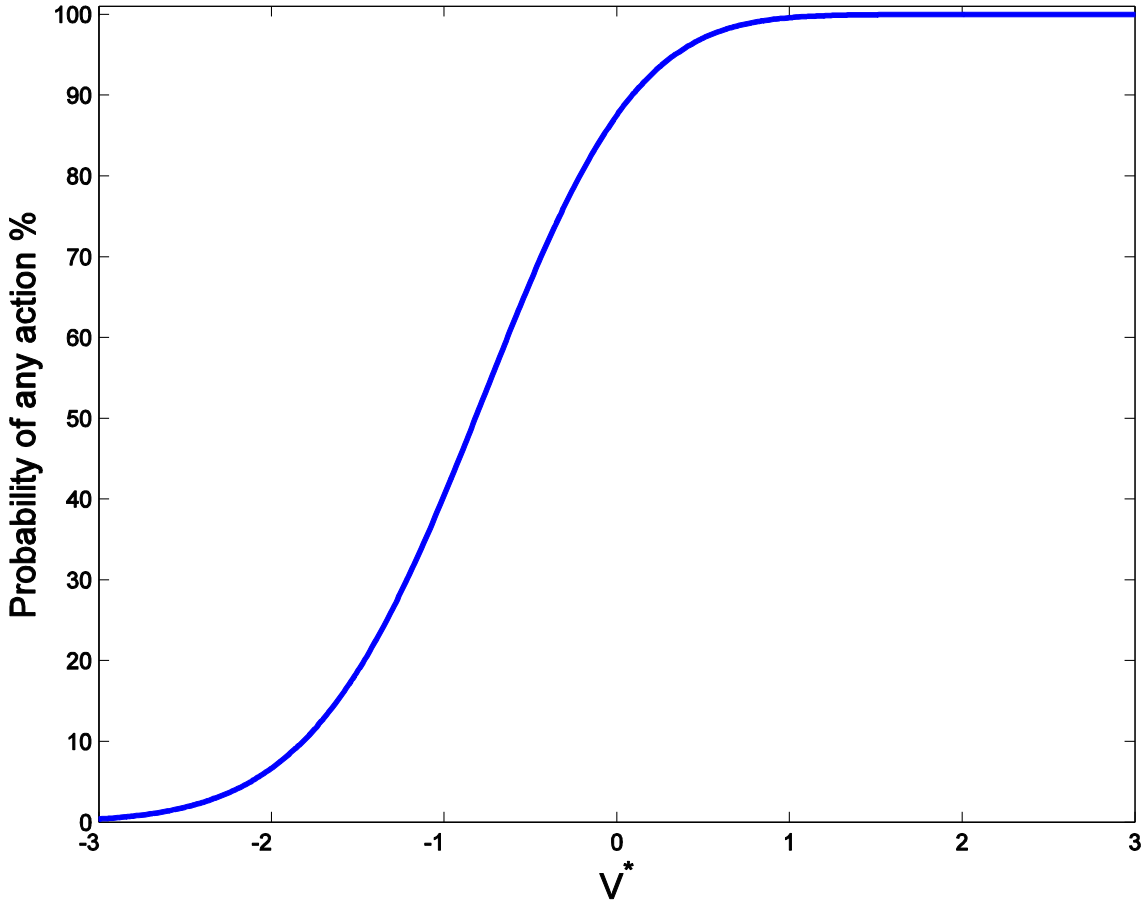

Figure 1: The probability of an initiative being undertaken as a function of the actual value, V, for 5 agents and assuming normally distributed errors with variance 1*

(these assumptions will be used in all subsequent figures except when otherwise noted). Note that 50% probability of action occurs near a value of -1: a strong unilateralist bias exists.

For mildly negative values of the initiative there is nearly always someone who misjudges the value of the initiative and undertakes it. There is no problem for positive initiatives since even if one or two agents are overly cautious, it is very likely that somebody will undertake the initiative, which is the optimal result.

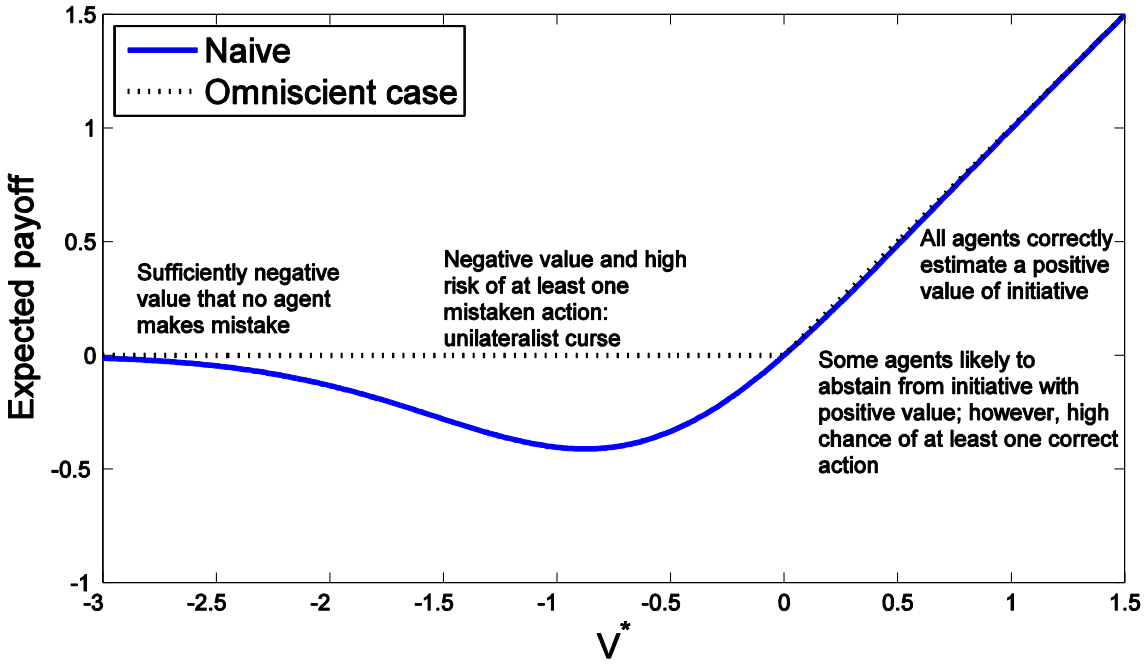

Figure 2: The expected payoff for naive agents (who act if and only if their evaluation of the initiative is positive) and ideal omniscient estimators who are assumed to know the true value.

Increasing the number of agents capable of undertaking the initiative also exacerbates the problem: as *N* grows, the likelihood of someone proceeding incorrectly increases monotonically towards 1.[7] The magnitude of this effect can be quite large even for relatively small number of agents. For example, with the same error assumptions as above, if the true value of the initiative $V^* = -1$ (the initiative is undesirable), then the probability of erroneously undertaking the initiative grows rapidly with *N*, passing 50% for just 4 agents.

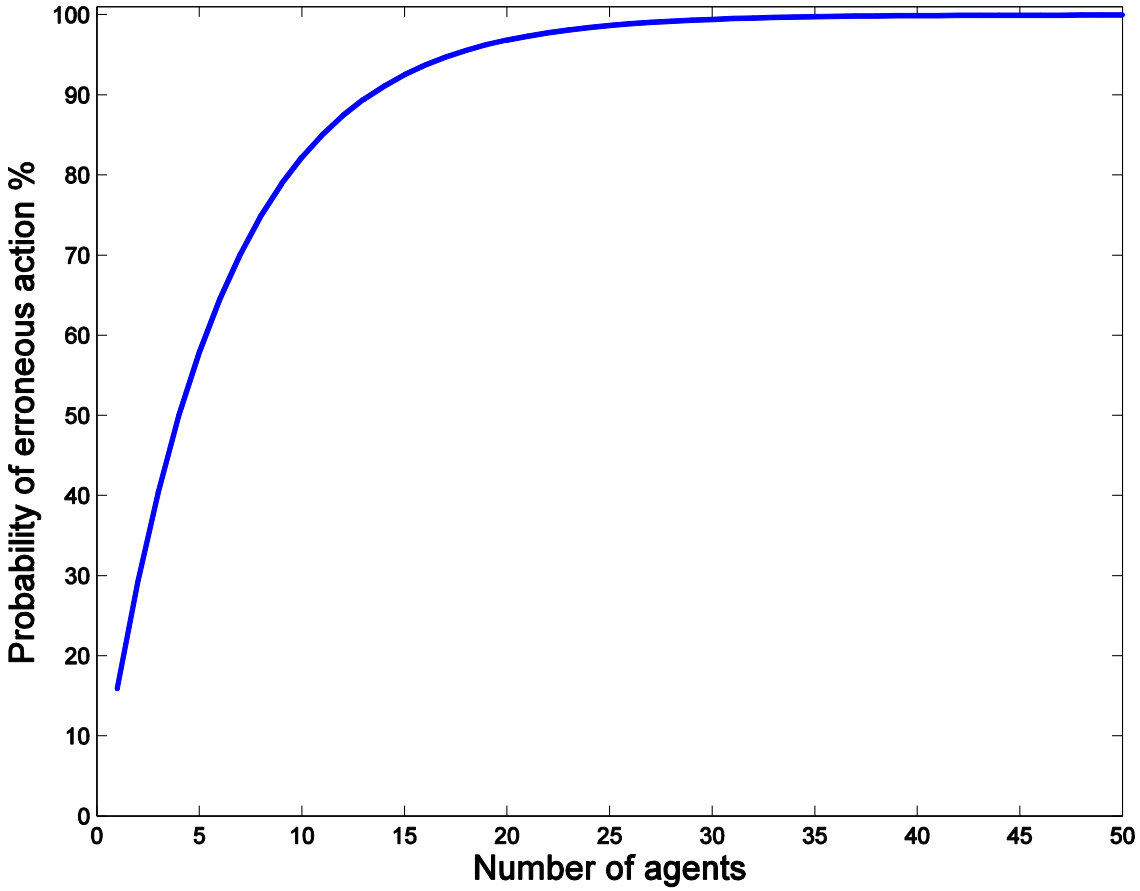

Figure 3: Probability of an erroneous action in the case of $V^* = -1$ for different numbers of agents.

There are six features of the unilateralist's curse that that need to be emphasized.

First, in cases where the curse arises, the risk of erroneously undertaking an initiative is not caused by self-interest. In the model, all agents act for the common good, they simply disagree about the contribution of the initiative to the common good.[8]

Second, though the curse could be described as a group-level bias in favor of undertaking initiatives, in does not arise from biases in the individual estimates of the value that would result from undertaking the initiative. The model above assumes symmetric random errors in the estimates of the true value.[9]

Third, there is a sense in which the unilateralist's curse is the obverse of Condorcet's jury theorem.[10] The jury theorem states that the *average* estimate of a group of people with above 50% likelihood of guessing correctly and with uncorrelated errors will tend to be close to the correct value, and will tend to move closer to the true value as the size of the group increases. But what is also true, and relevant to the argument in this paper, is that the *highest* estimate will tend to be above the true value, and the expected overestimation of this highest estimate *increases* with the size of the group. In the cases we are interested in here, it is the highest estimate that will determine whether an initiative is undertaken, not the average estimate.

Fourth, though we have chosen to illustrate the curse using initiatives that are (probably) irreversible, the problem can arise in other cases too. The curse becomes sharper if the initiative is irreversible, but even for actions that can be undone the problem remains in a milder form. Resources will be wasted on undoing erroneous initiatives, and if the bad consequences are not obvious they might occur before the problem is noticed. There might even be a costly tug-o-war between disagreeing agents.

Finally, fifth, though we have thus far focused on cases where a number of agents can undertake an initiative and it matters only whether at least one of them does so, a similar problem arises when any one of a group of agents can *spoil* an initiative—for instance, where universal action is required to bring about an intended outcome. Consider the following example:

> In Norse mythology, the goddess Hel of the underworld promised to release the universally beloved god Baldr if all objects, alive and dead, would shed a tear for him. All did, except the giantess Þökk. The god was forced to remain in the underworld.[11]

Similar situations can arise when all the actors in a play must come together in order for a rehearsal to take place, when all members of committee must attend a meeting in order for it to be quorate, or when all signatories to an international treaty must ratify it in order for it to come into effect. These cases are formally equivalent to the original unilateralist curse, with merely the sign reversed.

Since the problem in these cases is the result of *unilateral* abstinence, it seems appropriate to include them within the scope of the unilateralist's curse. Thus, in what follows, we assume that the unilateralist's curse can arise when each member of a group can unilaterally undertake *or spoil* an initiative (though for ease of exposition we sometimes mention only the former case).

3. Lifting the curse

Let a unilateralist situation be one in which each member of a group of agents can undertake or spoil an initiative regardless of the cooperation or opposition of other members of the group. We will say that a policy would lift the unilateralist's curse if universal adherence to it by all agents in unilateralist situations should be expected (*ex ante*) to eliminate any surfeit or deficit of initiatives that the unilateralist's curse might otherwise produce.

> *The Principle of Conformity*
> When acting out of concern for the common good in a unilateralist situation, reduce your likelihood of unilaterally undertaking or spoiling the initiative to a level that *ex ante* would be expected to lift the curse.

In the following subsections we will explore various ways in which one might bring oneself into compliance with this principle. These can be organized around three models: collective deliberation, epistemic deference, and moral deference. The three models are applicable in somewhat different circumstances, and their suitability might depend on the type of agents involved.

In addition to adhering to the principle of conformity in particular unilateralist situations, one might also have some moral reason to work at a more general level to counteract the unilateralist's curse. One way to do this would be to promote awareness and adoption of the principle of conformity. Another way would be to promote the development of institutions that make unilateralist situations less likely to arise, especially in regards to matters of global significance where the effects of the curse can be particularly devastating.

3.1. The collective deliberation model

A first line of defense against the unilateralist's curse could be to share data and reasoning between agents in the hope that this will resolve their disagreement about the desirability of proceeding with the contested initiative. Fully shared information is ideal, when it is achievable.

In some cases, however, extensive information sharing among all potential decision-making agents is impractical. Communication is often costly and time-consuming. Participants in a unilateralist situation may not even know of each other's existence. Furthermore, in certain cases information disclosure might itself be the initiative whose desirability is in dispute, such as when information hazards are associated with disseminating relevant data.[12]

Even when information is fully shared, a consensus can remain elusive. Disagreements about the net value of undertaking some project often persist after decision-makers have been thoroughly briefed on all obviously relevant and easily communicable facts and after having had opportunities to engage in joint deliberation.

Because complete information sharing may not be practical and because it may not produce consensus when it does occur, the principle of conformity requires us to explore additional models for lifting the unilateralist's curse.

3.2. The meta-rationality model

One approach would be to appeal to each agent's reflective rationality. A party to an epistemic disagreement should ideally reflect on the fallibility of their own judgment and adjust their posterior probability to take into account the fact that other agents have different opinions.

Robert Aumann has shown that rational Bayesian agents with identical priors and common knowledge of each other's posteriors (and of each other's rationality) must have identical posterior probabilities.[13] Disagreement between such agents is impossible. This sounds like good news: if all agents make the same estimate of the benefits of action, the unilateralist curse is lifted.

There is, however, some skepticism about the relevance of Aumann's result for practical cases of disagreement.[14] The assumption of identical priors, in particular, is problematic.[15] Furthermore, the same challenges that can make data sharing difficult can also make it difficult to make each agent's honest posterior probability estimates of the value of the initiative common knowledge among all agents.

It turns out, however, that sufficiently rational agents can manage the curse even without communication. In the literature on the winner's curse it has been argued

that rational expected utility-maximizing will not be affected by it.[16] Rational agents will take the winner's curse into account and adjust their bids accordingly. This is known as *bid shading*. Rational agents place bids that are lower than their *ex ante* expectation of the value of the good, but equal to their expectation of the value of the good conditional upon them winning the auction.

The counterpart in this response would be for agents in a unilateralist situation to estimate the value of the initiative conditional on the agent's first-order estimate of the initiative's value being the highest (or, in spoiler cases, the lowest).

In other words, on finding themselves in a unilateralist situation, each rational agent will initially estimate the value of the initiative based on his prior probability distribution. He will then take into account the case where his decision is decisive. In the case where agents can unilaterally undertake an initiative, the agent will condition on the situation in which he is the most sanguine and everybody else thinks the action should not be done. (In spoiler cases, the agent conditions on the situation in which he is the most pessimistic and everybody else thinks the initiative should be undertaken.) He then creates a posterior distribution of value that is used to make an adjusted decision.

$P(V^* \mid \text{win}) = P(\text{win} \mid V^*)\, P(V^*) / P(\text{win})$

Where "win" represents being the deciding agent.

> *Example:*
> In the simple case where the agent assumes all other agents have the same priors and are acting independently, only differing in the noisy data about V^* they have received,
>
> $$P(win|V^*) = \int_{-\infty}^{\infty} P(V - V^*)F(V - V^*)^{N-1}dV$$
>
> where *F(V)* is the cumulative distribution function of the errors. The posterior distribution of V^* becomes
>
> $$P(V^*|win) = KP(V^*)\int_{-\infty}^{\infty} P(V - V^*)F(V - V^*)^{N-1}dV$$
>
> where *K* is a normalization constant. The posterior action should then be based on the expectation $E(V^*|\text{win})$.
>
> If the agents choose to act when the received data is above a fixed threshold *T*, V^* is normally distributed with zero mean and variance 1, and they get estimates of V^* with normal noise (again with mean zero and variance 1), then the optimal threshold is the one that maximizes the expected value:
>
> $$T_{opt}(N) = argmax_T \int_{-\infty}^{\infty} VP(V)\left(1 - \left(1 - F(V - T)\right)^N\right)dV$$

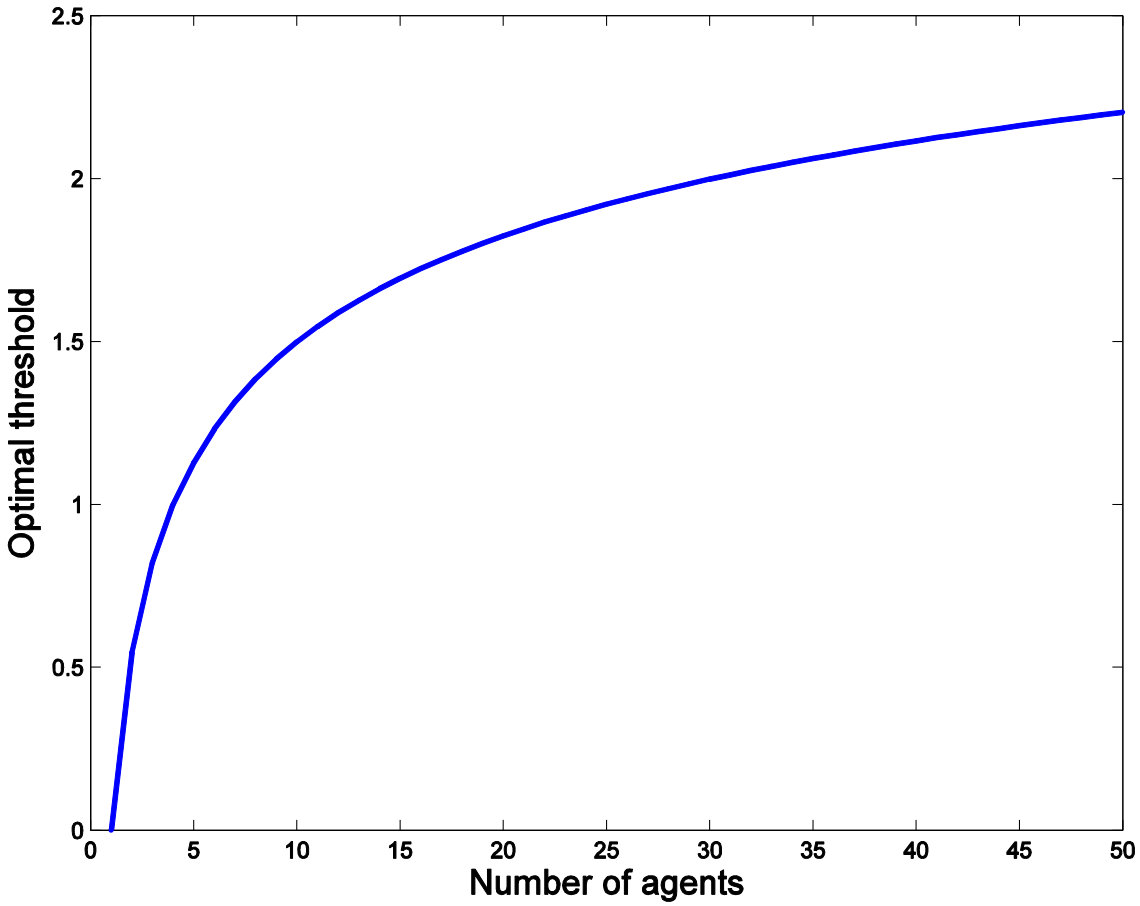

Figure 4: The optimal threshold $T_{opt}(N)$ for action as a function of the number of agents. Agents that only act if the perceived value of the initiative is higher than $T_{opt}(N)$ will maximize their expected (joint) result.

$T_{opt}(N)$ increases rapidly with N, reaching 0.54 for two agents and 1 for 4 agents: even for a small group it is rational to be far more cautious than in the single agent case. Note that in this case all agents are aware of the prior distribution, noise distribution, independence, and that the other agents are using this strategy.[17]

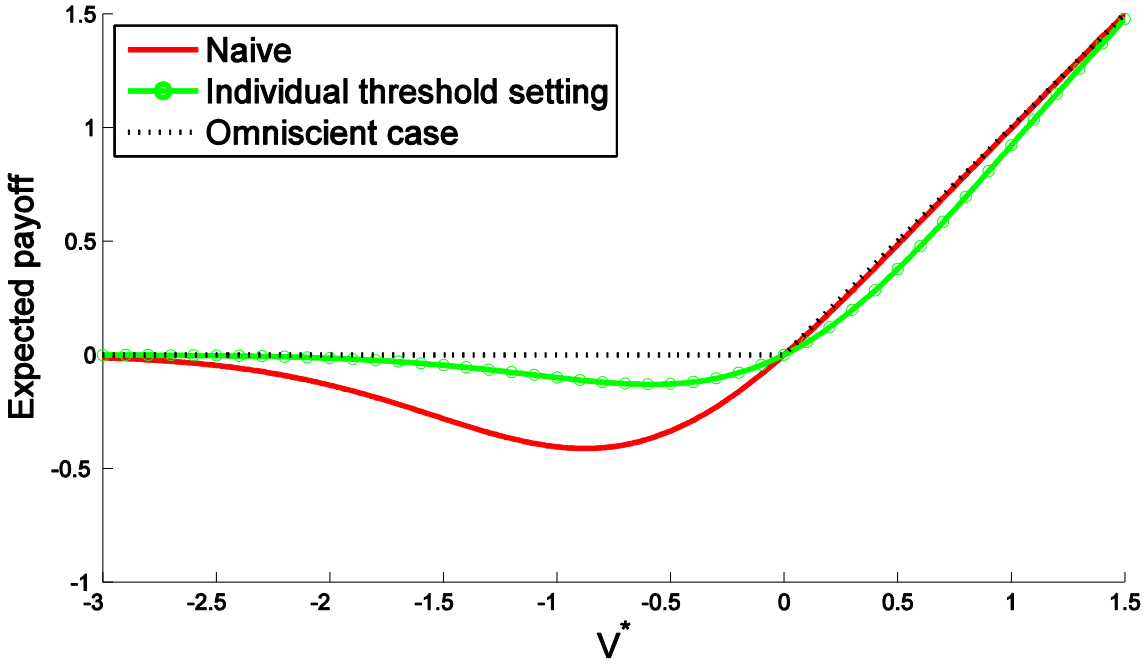

Figure 5: The expected payoff for different actual values of the initiative for alternative ways of handling the unilateralist's curse. Using the optimal individual threshold $T_{opt}(5)$ reduces the losses significantly.

One might raise questions about the practical applicability of this sophisticated Bayesian approach, however. Even if rational Bayesian agents would agree, humans are at best approximations of rational Bayesian agents and they have far more limited mental computation power—even when leaving out biasing factors.[18] Value in practical cases is also seldom in the form of easily manipulable and comparable scalar quantities. Hence implementing the sophisticated Bayesian approach to lifting the unilateralist's curse might typically be infeasible.[19]

3.3. The moral deference model

Suppose a unilateralist situation exists and that it is not feasible for all agents to lift the curse through communication and adjustment of beliefs. It might nevertheless be possible for the group to lift the curse if each agent complies with a moral norm which reduces the likelihood that he acts unilaterally, for example, by assigning decision-making authority to the group as a whole or to one individual within it. We call this the moral deference model.

In contrast to the two models presented above, the moral deference model does not require agents to defer to the group in forming their beliefs regarding the value of the initiative. However, it does require them to defer to the group in deciding whether to act on those beliefs. A slogan for this approach could be 'comply in action, defy in thought'.

There are many norms such that universal compliance with the norm by a group of agents would lift the unilateralist's curse. For example, a norm that assigned decision-making authority to an arbitrary member of the group would lift it. Consider the norm: when in a unilateralist situation, if you are the tallest person able to

undertake the initiative, then undertake it if and only if you believe its value exceeds zero; if you are not the tallest person able to undertake the initiative, do not undertake it.

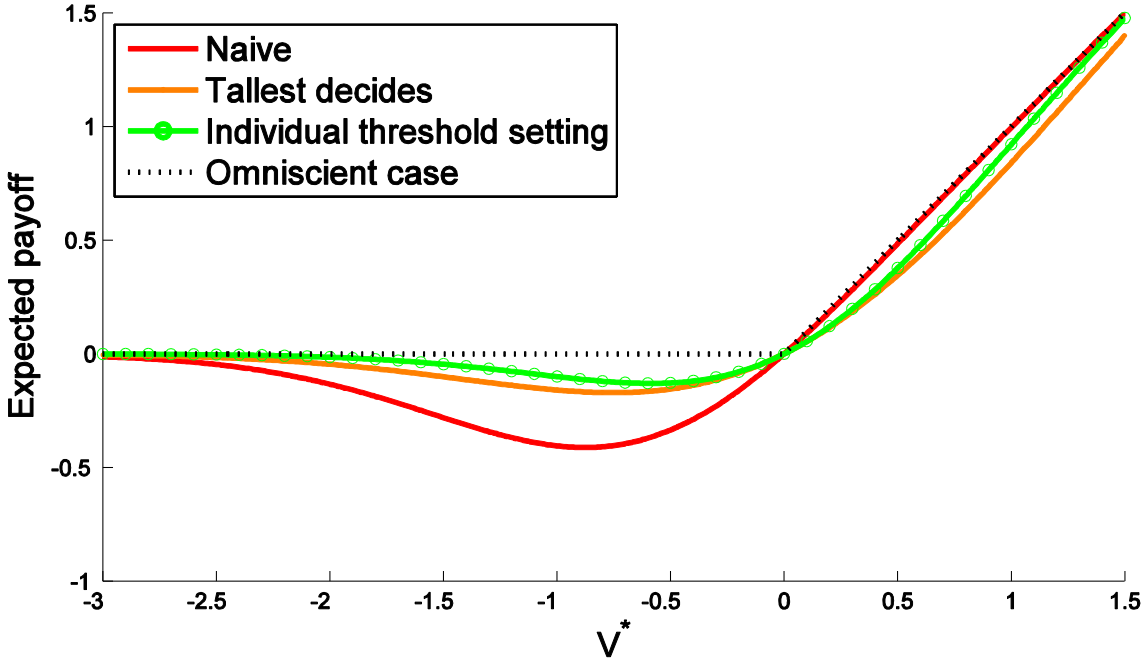

Figure 6: Expected payoff for different actual values of the initiative for alternative ways of handling the unilateralist curse. The tallest decides case achieves a significant reduction of loss, nearly reaching the payoff of the more complex Bayesian threshold method.

Universal compliance with this norm would prevent the unilateralist's curse from arising in the sense that, in the absence of any bias towards or against action in the individual members of the group (and thus in the group's tallest member), this norm will produce no group-level bias towards or against the initiative.[20] The payoffs associated with this tallest-decides norm in a five-agent situation are depicted in figure 6 above. The tallest-decides norm, however, has several unattractive features. For example, it does not protect against biases or errors that might impair the judgment of the group's tallest member. Furthermore, it is very unlikely that such a norm would gain wide acceptance.

Fortunately, there are other norms that could lift the curse and lack these unattractive features. One appealing norm would recommend that agents conform to the rules of existing institutions that militate against unilateral action:

> (1) When in a unilateralist's situation, defer to existing institutions, such as laws or customs, if universal deference to those institutions would lift the unilateralist's curse.

National and international laws often militate against the unilateralist's curse, for example by specifying that decisions must be made democratically or by individuals or institutions that have been given special authority over a particular realm of decision-making. In other cases, there are informal conventions that may do the job.

For example, following the publication early last decade of two studies thought by some to aid bioweapons development,[21] a group of scientific journals agreed to introduce screening procedures to identify papers containing information that is especially prone to misuse and to seek external advice on the publication of such papers.[22] Though these procedures lacked legal status, compliance with them by journals may have helped lift the curse.

One virtue of (1) is that, since it simply reinforces existing institutional norms which may already command significant support, it may be relatively easy for it to achieve wide acceptance. However, (1) will not lift the curse in all cases. In many areas with an international dimension, for example, there are no relevant international laws and deference to national laws would merely create a new unilateralist situation between nations: the nation that evaluates the initiative most positively is most likely to allow it.

It might be possible for a group of agents to lift the curse even in cases where (1) fails by complying with a different norm, one that promotes the development of and compliance with a new procedure for group decision-making. For example, suppose all agents faced with a unilateralist situation complied with the norm:

> (2) When in a unilateralist's situation, promote the holding of a majority vote among those capable of undertaking the initiative. If the vote takes place, then (a) defer to its verdict, and (b) encourage others to do likewise.

Universal compliance with this norm is likely to lift the curse. Since it is effectively using the median estimate it is robust to outliers. It will also tend to reduce systematic bias at the group level provided that individual biases are at least partially independent of one another.[23] And since majority voting is a common and widely accepted method for group decision-making, this norm would have relatively good prospects of gaining wide acceptance.

Compliance with norms (1) and (2) will, however, lift the unilateralist's curse only when a high degree of communication and coordination is possible. There are other norms whose universal adoption could lift the curse even in the absence of communication and coordination. Consider the norm:

> (3) When in a unilateralist situation, bring about the outcome if and only if you judge that a majority vote among those capable of undertaking the initiative would yield a majority in favor of doing so.

Insofar as each individual capable of undertaking the initiative makes an accurate prediction of the views of all others, universal adoption of this norm will eliminate any group-level bias due to the unilateralist's curse. Even if predictions of the views of others are inaccurate (for example, because each agent overestimates the extent to which others share her views), universal adoption of this principle can still be expected to somewhat mitigate the unilateralist's curse. It will tend to reduce the likelihood of undertaking an initiative of those who assess the value of the initiative most favorably, provided that these agents realize they are at the optimistic end of the spectrum.[24]

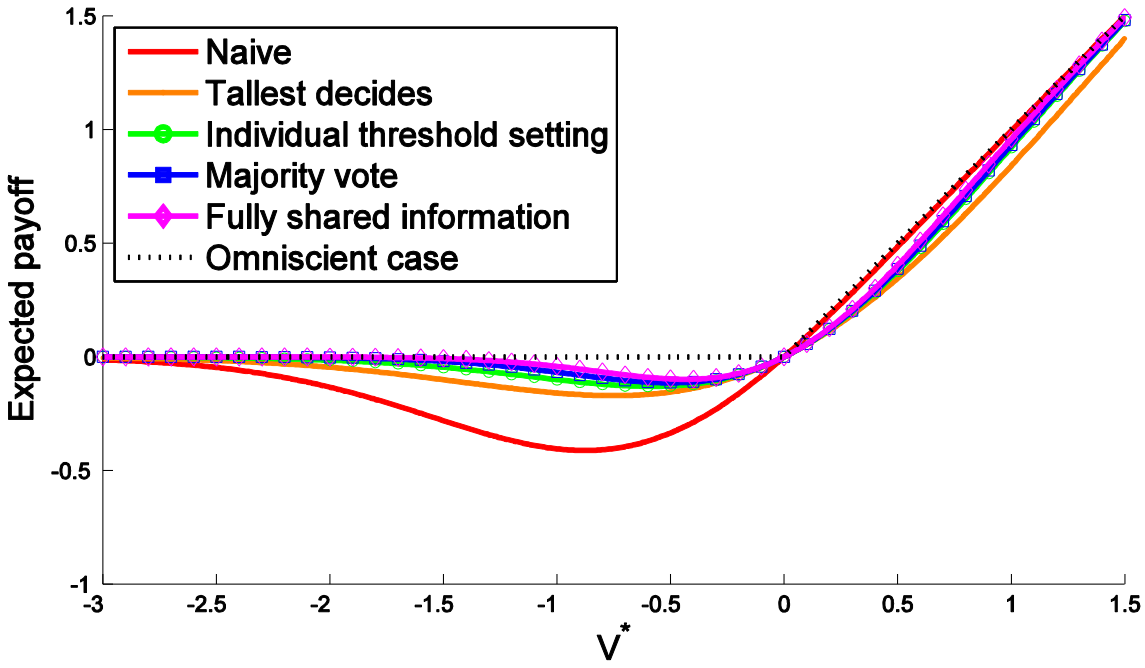

Figure 7: The expected payoff associated with universal compliance with six different strategies at different actual values of the initiative. The fully shared information strategy consists in pooling the information between the agents and acting on the group's best joint estimate of V^*;[25] this requires maximal communication. Despite the lack of communication in tallest decides and threshold setting, the agents achieve an average outcome close to the cases where communication is possible.

Figure 7 depicts, for a five-agent case, the expected payoffs associated with two of the norms discussed in this section—tallest decides, and the actual majority vote (norm (2))—and it compares these with other strategies described in section 3.2 above. Under our assumptions, the majority vote does rather well—it is close to the maximum available payoff represented by the omniscient case.

However, in the real world, different strategies will work well in different cases. It is thus likely that the best norm to adopt, under the moral deference model, would be some composite of simple norms such as (1)-(3). For example, a group might adopt a norm that specifies that the group should act as specified by (1), (2) or (3) depending on what laws and conventions already exist, what forms of communication and coordination among group members are possible, and how costly such communication and coordination is likely to be, among other factors.

We do not wish to commit ourselves to norms (1)-(3) as the best building blocks from which to construct such a composite norm. We believe that each of (1)-(3) is at least a plausible candidate for inclusion in a composite norm. However, there may be other norms that would more fully lift the curse or which have other advantages over (1)-(3). For example, there are well-known problems with majority voting which should perhaps lead us to prefer a different voting procedure under norms (2) and (3).

One other set of concerns regarding norms (2) and (3) warrants mentioning. Both of these norms involve holding a vote (real or hypothetical) *among agents capable of undertaking the initiative in question*. But it might be argued that any actual or hypothetical vote should include more individuals than merely those capable of undertaking the initiative. It could be argued, for example, that the vote should include all individuals who will be affected by the initiative. Consider a case in which there are three agents who could undertake an initiative and two of the three judge that it would be best to do so. However, millions of others will be affected by the initiative and almost all of them judge that the initiative has net disvalue. In this case, it might seem odd that a vote among the three agent's capable of undertaking the initiative should be preferred over a vote among all who would be affected by it.

A more specific problem with excluding individuals who are incapable of undertaking the initiative is that this might seem to skew the vote. There might be some agents who are not capable of undertaking the initiative, but could have been capable of doing so; they are incapable only because they previously judged that undertaking the initiative would be a bad idea and thus ceased to develop the necessary capacities. Excluding these agents from a vote might seem to skew the vote in favor of those who deem the initiative to be valuable and who have thus sought to develop the capacities necessary to undertake it.[26]

At the same time, it might be argued that some agents capable of undertaking the initiative should be *excluded* from the vote. Suppose that each of five nations is capable of undertaking some geoengineering project with worldwide consequences. Four agree to hold a majority vote among the five nations and to abide by the outcome of that vote. The fifth wishes to take part in the vote but is resolved to press ahead with the project regardless of the outcome of the vote. It might seem doubtful whether the first four nations should include the fifth in the vote. Arguably, deferring to a majority vote in unilateralist cases involves making a sacrifice. It involves giving away some of one's autonomous decision-making authority. It might seem that it would be unfair for the fifth nation to exert an influence over the decisions of others by participating in a vote without also being prepared to make the same sacrifice that the others are prepared to make. This may count in favor of excluding the fifth nation. Excluding the fifth nation might also help to incentivize deference to majority votes in unilateralist situations.

There are thus arguments both for expanding and for restricting the group of agents given a vote in norms (2) and (3). We cannot assess these arguments here. We mention them only to flag them as topics for further discussion. However, it is worth noting that including all and only those agents who are capable of undertaking an initiative does at least have the virtue of picking out a group that would, in many cases, be relatively easy to identify.

We should end this section on the moral deference model with an important clarification: the model does not rely on a commitment to any particular moral theory. Proponents of a range of different moral theories could accept norms of the sort described above, though they would assign different statuses to them.

A rule consequentialist, for example, might treat these norms as genuine moral principles—principles that determine which acts are right and which are wrong. According to one formulation of rule consequentialism, a rule of action is a genuine moral principle just in case it is part of the set of rules of action whose general

acceptance can be expected to have consequences as good as the general acceptance of any alternative set of rules.[27] Given the risk of premature or erroneous action created by the unilateralist's curse and the likelihood that most agents are not sophisticated enough belief-formers to apply our meta-rationality model, it is plausible that the optimal set of rules will contain a norm of the sort that we have discussed.

On some other moral theories, these norms would serve not as genuine moral principles, but as guidelines for helping agents to comply with such principles. Adherents of many moral theories, both consequentialist and deontological, could accept something like the following moral principle:

> Agents have moral reasons to undertake an initiative if and only if that initiative would contribute to the common good, and to spoil an initiative if and only if that initiative would detract from the common good.

Norms of the sort discussed above could help agents to better comply with this principle in unilateralist situations.[28]

4. Discussion

We proposed:

> *The Principle of Conformity*
> When acting out of concern for the common good in a unilateralist situation, reduce your likelihood of unilaterally undertaking or spoiling the initiative to a level that *ex ante* would be expected to lift the curse.

We also outlined three different ways in which agents who find themselves in unilateralist situations might comply with this principle. We do not claim that any one of these models is superior to the others in all situations. Which model should be adopted will depend, among other things, on the sophistication of the agents, the degree of communication and coordination that is possible, and the nature of existing laws and conventions bearing on the decision.

In this section we discuss two concerns that might be raised regarding our principle.

4.1. The historical record

Adoption of the principle of conformity is meant to make things better. Yet if we 'backtest' the principle on historical experience, it is not at all clear that universal adoption of the principle of conformity would have had a net positive effect. It seems that, quite often, what is now widely recognized as important progress was instigated by the unilateral actions of mavericks, dissidents, and visionaries, who undertook initiatives that most of their contemporaries would have viewed with hostility and that existing institutions sought to suppress. The benefits of iconoclasm and defiance of authority have been stated especially forcefully in the Enlightenment tradition and by proponents of scientific and technological progress:

> 'Every great advance in natural knowledge has involved the absolute rejection of authority.'
>
> — *Thomas Huxley*

> 'There is no great invention, from fire to flying, which has not been hailed as an insult to some god'
>
> *— J. S. B. Haldane*

The principle of conformity could be seen to imply, for instance, that Galileo Galilei ought to have heeded the admonitions of the Catholic Church and ceased his efforts to investigate and promote the heliocentric theory.[29] (Similarly awkward implications would hold for various religious initiatives that were unpopular at the time.) It is embarrassing for our argumentation that it appears to have such implications.

It is possible that the appearance that unilateralism has historically been mostly for the good is illusory. Historical unilateralism might be more salient when it worked out well than when it worked out badly, perhaps because successes have been more extreme but less frequent than the failures.

However, even if unilateralism *has* historically provided a net benefit to humanity, this need not undermine our argument. The claim that the unilateralist curse is an important phenomenon and that we have reason to lift it is consistent with the claim that the curse has provided a net benefit to humanity.

One way that the two claims can be reconciled is by noting that the main effect of the curse is to produce a tendency towards unilateral initiatives, and that if it has historically been the case that there have been other factors that have tended to strongly inhibit unilateral initiatives, then it could be the case that the curse has had the net effect of moving the overall amount of unilateralism closer to the optimal level. For example, it might be argued that the scholars of past ages were usually far too deferential to authority, for reasons independent of the factors discussed in this paper. Their failure to take into account our arguments might then have had the salutary effect of not further inhibiting whatever propensity remained to promote new thoughts.

Another way that the two claims might be reconciled is by invoking luck. Even if it were the case throughout most of history that, on net, key actors would have moved closer to the rationally preferable level of unilateralism if they had adopted the principle of conformity, it does not follow that consequences of this increased rationality would have had to be positive. It is possible for irrationality to pay off. A gambler might feel irrationally confident in the outcome of a dice roll, and if he is lucky he might benefit from his irrationality.

It might be objected that human history has been going on for a while and that it would be highly improbable that an irrational betting strategy would have kept paying off in so many instances; the long run favors the house. Therefore, it might be said, the invocation of luck would be a very feeble defense of our view. But this is not necessarily so if the outcomes of the individual bets are correlated: in that case, there might have in effect been a single bet, albeit one in which many different people have participated on many different occasions—a single net long-position on the benefits of science and innovation. Even if this were irrational (as judged by the information available at the time) it need not be that surprising that it should have proved a winner. And the failure to heed the reasoning for the principle of conformity could be viewed as having resulted in just such a long-position.

These two ways of reconciling our argument with the hypothesis that it would have been on net bad if the principle of conformity had been widely adopted historically carry different lessons for prospective decision-making. If the past discrepancy is due to luck, then there is no reason not to simply embrace the principle and adopt a more conforming stance in unilateralist situations. If, in contrast, the historical discrepancy is due to the curse serving to counteract some other factor that biases action towards conformity, then it is not clear that lifting the unilateralist's curse is desirable. Lifting the curse would still be desirable if the countervailing historical biases have now ceased to operate (for instance, if there is no longer a strong tendency towards herd mentality or towards deferring excessively to authority). If, however, some of these biases are still in effect, then simply removing the unilateralist's curse could make things worse.

One might instead attempt a more complicated intervention aiming at the simultaneous removal of the biases of conformism and the biases of unilateralism. If this could be done, it should have a generally clarifying effect on our thinking and place our individual and public deliberations on a sounder and securer foundation. If it could not be done, then introducing our principle of conformity might be the equivalent of putting on a nice new boot on the left foot while retaining an old moccasin on the right: a local improvement that makes things worse overall.

4.2. Empowering powerful groups

Another way in which adoption of the principle of conformity might make things worse is by preferentially boosting the coordination ability of groups that are already powerful and able to undertake many kinds of initiative. If one believes that it is bad for the world to increase power-differentials by making such already powerful groups more powerful, then one might disfavor changes that make it easier for such groups to coordinate internally to attain their aims. Consider, for example, a set of powerful mafia bosses who have joined to form a criminal cartel. Each of them has the ability, by withdrawing his cooperation, to destroy the cartel. In a case like this, society may be better off if the bosses find it harder to cooperate—for example, if each of them is disposed to withdraw from the cartel as soon as it seems to him best to do so, independently of what his peers think.[30]

Or consider the case of a whistle-blower like Daniel Ellsberg, famous for leaking the Pentagon Papers. Most of Ellsberg's peers, who had the high-level security clearance required to access the relevant documents, presumably did not believe that leaking the material to the press would contribute positively to the common good. If Ellsberg had sought to follow the principle of conformity, for example by imagining a vote among all those in a position to leak the documents, it would seem he would have had to conclude that the documents ought not be leaked. Those who might have had the most positive evaluation of the information disclosure (such as the American public or opponents to the Vietnam war around the world) were not in a position to undertake that initiative.

Cases like these illustrate the importance of a point we made earlier: it makes a difference how the group of (imaginary or actual) voters or epistemic peers is defined. If one allows that these groups might be defined more broadly than the group of agents capable of undertaking an action, it may be possible to avoid the unpalatable implication that Ellsberg should have refrained from whistleblowing. Perhaps many 'outsiders' would have (hypothetically) voted in favor of his release of information.

5. Concluding thoughts

We have described a moral analogue of the winner's curse. The unilateralist's curse arises when each of a group of agents can, regardless of the opposition of others, undertake or spoil an initiative that has significant effects on others. In such cases, if each agent decides whether to undertake (or spoil) the initiative based on his own independent naive assessment of its value, there will be a group-level bias towards undertaking (spoiling) the initiative. Importantly, this effect arises even if all the agents are assumed to be motivated solely by concern for the common good.

We proposed a principle—the principle of conformity—which instructs agents faced with a unilateralist situation to reduce their likelihood of unilaterally undertaking (or spoiling) the initiative. We then outlined three models for accomplishing this. They involved, respectively, (1) sharing information and reasoning before forming one's evaluation of the initiative, (2) adjusting one's evaluation in the light of the curse, and (3) deferring to the group in making one's decision.

As we acknowledged in the previous section, there may be considerations that militate against the principle of conformity. For example, if there is already a group-level bias against unilateralism, then compliance with the principle would exacerbate this bias. However, we maintain that there is a *prima facie* case for complying with the principle. Moreover, since the level of bias due to such other factors towards or against unilateralism presumably varies across different contexts, it is likely that there will be some contexts in which the *prima facie* case for complying with the principle will be decisive. Those will be the contexts in which the group-level bias due to the unilateralist's curse is greater than the any countervailing bias against unilateralism.

It is also possible that, at least within the domain of science, the principle of conformity is more relevant today than it was in, say, Galileo's time. At that time, there was, plausibly, a strong bias against thinking and acting independently in intellectual matters, at least where this would involve diverging from the views of the Church. Since the Enlightenment, however, there may have been a significant weakening of this bias. Independence of thought and action is now more widely regarded as a virtue in scientists and other intellectuals. Honors and prizes are won based on claims to originality and precedence. There may now be no bias, or only a weak bias, against unilateralism in science. Thus, the risk posed by the unilateralist curse in scientific contexts may be greater now than ever.

To resist the unilateralists' curse one first has to become aware of when one is in a curse situation. We hope this paper will help achieve that.

Acknowledgements

[Removed for blind review.]

References

Armstrong, S. (2012): "Nash equilibrium of identical agents facing the Unilateralist's Curse", Technical Report #2012-3, Future of Humanity Institute, Oxford University: pp. 1-5.

Atlas, Ronald, Philip Campbell, Nicholas R Cozzarelli, Greg Curfman, Lynn Enquist, Gerald Fink, Annette Flanagin, et al. "Statement on Scientific Publication and Security." *Science (New York, N.Y.)* 299, no. 5610 (February 21, 2003): 1149.

Aumann, Robert J. (1976). "Agreeing to Disagree". *The Annals of Statistics* 4 (6): 1236–1239.

BBSRC, MRC, and Wellcome Trust. 2005. Managing risks of misuse associated with grant funding activities: A joint BBSRC, MRC and Wellcome Trust policy statement. Available at http://www.bbsrc.ac.uk/organisation/policies/position/public_interest/misuse_of_research_joint.pdf (accessed Jan 14, 2009).

Nick Bostrom, "Information Hazards: A Typology of Potential Harms from Knowledge". *Review of Contemporary Philosophy*, Vol. 10 (2011): pp. 44-79 (http://www.nickbostrom.com/information-hazards.pdf)

Bowden, C. "Our Wall." National Geographic 211 (2007): 115–139.
Christensen, D. (2009). "Disagreement as Evidence: The Epistemology of Controversy", *Philosophy Compass*, 4/5: 756-767.

Cello, J., A. V. Paul, and E. Wimmer. 2002. Chemical synthesis of poliovirus cDNA: Generation of infectious virus in the absence of natural template. *Science* 297(5583): 1016-18.

Marquis de Condorcet. Essai sur l'application de l'analyse á la probabilité des décisions rendues á la pluralité des voix, 1785.

Cox, James C. and Isaac, R. Mark, In search of the winner's curse. *Economic Inquiry*, 22:4, p. 579-592 1984

Cowen, Tyler and Hanson, Robin, Disagreement as Self-Deception About Meta-Rationality, 2001 http://holtz.org/Library/Philosophy/Epistemology/Disagreement%20as%20Self-Deception%20About%20Meta-Rationality%20-%20Hanson%202002.pdf

Michael Davis, Avoiding the tragedy of whistleblowing, *Business & professional ethics journal*, vol 8: 4, 1989 p. 3-19.

Faden, Ruth R, and Ruth A Karron. "The Obligation to Prevent the Next Dual-Use Controversy." *Science* 335, no. 6070 (February 17, 2012): 802–804.

Feldman, R. & Warfield, T. (eds.). *Disagreement* (Oxford: Oxford University Press, 2010).

Grofman, Bernard; Guillermo Owen; and Scott L. Feld. 1983. Thirteen theorems in search of the truth. *Theory & Decision*, 15: 261-78.

Hanson, Robin (2006). "Uncommon Priors Require Origin Disputes". *Theory and Decision* 61 (4): 319–328.

Hooker, B. *Ideal Code, Real World: A Rule-Consequentialist Theory of Morality*. Oxford: Clarendon Press, 2002.

Jackson, R. J., A. J. Ramsay, C. D. Christensen, S. Beaton, D. F. Hall, I. A. Ramshaw. 2001. Expression of mouse interleukin-4 by a recombinant ectromelia virus suppresses cytolytic lymphocyte responses and overcomes genetic resistance to mousepox. *Journal of Virology* 75(3): 1205-1210.

Journal Editors and Authors Group. 2003. Uncensored exchange of scientific results. *Proceedings of the National Academy of Sciences* 100(4): 1464.

Ladha, Krishna K. The Condorcet Jury Theorem, Free Speech, and Correlated Votes. *American Journal of Political Science*, Vol. 36, No. 3 (Aug., 1992), pp. 617-634

Osterholm, Michael T., and Donald A. Henderson. "Life Sciences at a Crossroads: Respiratory Transmissible H5N1." *Science* 335, no. 6070 (February 17, 2012): 801–802.

The Progressive Magazine. "The H-Bomb Secret: How We Got It and Why We're Telling It." The Progressive Magazine, November 1979 (full issue).

Perez, Daniel R. "H5N1 Debates: Hung Up on the Wrong Questions." *Science* 335, no. 6070 (February 17, 2012): 799–801.

Joseph Raz, 'Authority, Law, and Morality' in his, *Ethics in the Public Domain* (Oxford: Clarendon Press, 1994).

Royal Society, *Geoengineering the climate: science, governance and uncertainty*, RS Policy document 10/09, The Royal Society, September 2009

Schneier, Bruce. Locks and full disclosure. *IEEE Security and Privacy*, 1:2, March 2003

Richard H. Thaler, Anomalies: the winner's curse, The Journal of Economic Perspectives, vol 2:1, pp. 191-202, 1988.

Victor D (2008). On the regulation of geoengineering. *Oxford Review of Economic Policy* 24, 2, 322–336.

Williams, K., Australia Bureau of Resource Sciences, CSIRO Division of Wildlife, and Ecology. *Managing Vertebrate Pests: Rabbits*. Australian Government Publishing Service Canberra, 1995.

Eckard Wimmer, The test-tube synthesis of a chemical called poliovirus: The simple synthesis of a virus has far-reaching societal implications, E*MBO Rep*. 2006 July; 7(SI): S3–S9.

---

[1] We assume that the common good is determined in part by the wellbeing of all persons and other morally significant individuals. However, we make remain neutral on precisely how individual wellbeing determines the common good. For example, we do not commit ourselves to the view that the common good is simply aggregate individual wellbeing; we allow that the distribution of wellbeing might be relevant. We also allow that factors besides individual wellbeing might influence the common good. For example, some initiatives might possess intrinsic value that is independent of their contribution to wellbeing, and we allow that this intrinsic value might be one element in the common good.

[2] (The Progressive Magazine, 1979)

[3] (Bowden, 2007)

[4] (Williams, 1995)

[5] (Thaler, 1988)

[6] The probability that a particular agent will be wrong about the sign of the value of the outcome is $Pr(V^* + d > 0)$ if $V^*<0$ and $Pr(V^* + d < 0)$ if $V^*>0$. This is equal to $1-F(-V^*)$ if $V^*<0$ and $F(-V^*)$ if $V^*>0$. The probability that out of N agents at least one will be wrong about the sign is $(1-F(-V^*)^N)$ if $V^*<0$ and $(1-(1-F(-V^*))^N)$ if $V^*>0$. However, even if errors are symmetric around 0, the expected outcome is not: in the $V^*<0$ case it is enough that one agent acts for a negative value to be obtained, while in the $V^*>0$ case all agents have to err on the side of caution for them to lose out on a positive value. The expected value obtained by naive agents is hence $V^*(1-F(-V^*)^N)$. For positive values this is close to V^* (for unbiased error distributions), and we will hence focus on the $V^*<0$ case where unilateral action is a problem.

[7] Theorem: As N grows, the likelihood P of at least one agent proceeding incorrectly increases monotonically towards 1 unless $F(-V^*) = 1$ (i.e. unless there is an upper limit on the size of the deviations and V^* is more negative than this limit: no agent will ever make a sufficiently bad mistake).

Proof: If $F(-V^*) = 1$, $P = 0$ for all N. Otherwise $0 \leq F(-V^*) < 1$, and hence $F(-V^*)^N$ approaches 0 as $N \rightarrow \infty$.

[8] There will also, of course, be cases where an agent's decision whether to undertake an initiative affects others but the agents are motivated by self-interest rather than the common good. In these cases, there are two possible reasons for getting the wrong decision, from the point of view of the common good: (i) self-interest and the common good come apart—that is, one is judged to have positive value and the other negative value—and (ii) the agent overestimates 'self-interest' value.

[9] If the distribution of errors is *skewed* such that the typical estimate is higher than the true value, for instance due to optimism bias, then the risk of erroneous action is increased: in that case, even a single agent might be likely to overestimate the value of the initiative sufficiently to undertake it even when the true expectation value of the initiative is strongly negative. But this is unrelated to the curse.

In the case of estimates skewed towards safety—that is, there is pessimism bias—any tail distribution allowing mistaken action will still produce a growing probability of going ahead as N grows, although there may be intermediary cases where the curse would helpfully serve to balance out an opposite effect arising from pessimism bias. However, this situation may be rare.

[10] (Condorcet 1785)

[11] Cf. Snorri Sturluson's *Gylfagining*.

[12] (Bostrom 2011)

[13] (Aumann 1976)

[14] For discussion, see e.g. (Christensen 2009) and (Feldman & Warfield 2012).

[15] Attempts to weaken this assumption have been made; see (Hanson 2006).

[16] (Cox & Isaac 1984)

[17] In actual cases, the other agents are likely to have different priors and non-independent information, plus uncertainty about the number of agents. This possibility can be included in our the top equation, at the price of a far more complex model that needs several priors.

[18] Including self-deception about how meta-rational they are. (Cowen and Hanson 2001)

[19] Another way of looking at the problem is through the lens of game theory. Each agent needs to choose a (pure or mixed) strategy mapping their observations into actions, trying to maximize expected utility. We assume that all agents share a single utility function, i.e. they are all working for the common good. Since the agents know they are identical and will not be able to communicate, they will be using the same strategy. It can then be shown that there if there is any local maximum in their utility function if they all use the same strategy *g*, then the general use of *g* is a Nash equilibrium. (See (Armstrong 2012) for further details). The equilibrium can be non-strict under some conditions: a single agent is free to follow a different strategy without changing the outcome. This means that no agent will be able to realise higher expected value pursuing a different strategy.

Note that optimal strategies can be probabilistic (i.e., mixed). For example, suppose the information each agent received is either a red light or a green light (indicating whether the initiative should be undertaken), but the green light is only correct 75% of the time. For multiple agents, always undertaking the initiative when a green light is received produces a worse outcome than only acting on a green light with a probability less than one. As the number of agents goes up this probability should become lower, exploiting the fact that in the case the action does have positive outcome the likelihood of at least one agent acting remains high enough. Calculating the optimal probability requires an estimate of the number of agents and the probability of erroneous information, again requiring Bayesian priors. Game theory mainly tells us that a solution exists, but finding it requires the meta-rationality approach.

[20] The norm does not deal with 'spoiler' cases, where one agent can prevent an initiative from taking place. However, an analogous norm could be adopted to lift the unilateralist curse in those cases.

[21] Jackson et al. (2001); Cello et al. (2002).

[22] See Atlas et al. (2003), Journal Editors and Authors Group (2003). This procedure was invoked in the wake of two recent studies which which demonstrated how to make avian influenza transmissible by air between ferrets. See, for discussion, Perez (2012); Faden and Karron (2012); Osterholm & Henderson (2012).

[23] The assumptions of the Condorcet theorem can be weakened in many ways. In particular, agent competence only has to be on average above 50% (Grofman, Owen & Feld 1983), and a certain level of voting correlation does not reduce majority voting performance (Ladha 1992).

[24] For similar reasons, an analogous norm would tend to reduce the likelihood of *spoiling* an initiative of those who evaluation an initiative most negatively.

[25] In this case, the maximum likelihood estimate is simply the average of their individual estimates.

[26] One problem with including individuals incapable of undertaking the initiative in the majority vote is that these agents may lack information that bears on the value of the initiative, and it may be impossible or undesirable to provide them with this information. For example, suppose that the initiative under consideration is the release of dangerous scientific knowledge, such as knowledge about how to render HIV transmissible by air. Suppose only a few HIV experts have this information but that each could make it widely available through publication. In this case, outsiders lack a crucial piece of information relevant to the value of the initiative: they are ignorant of the content of the information whose release is in question. Moreover, it is not possible to give them this information without undertaking the initiative and rendering the evaluation moot.

[26] When the vote is merely hypothetical, there may be a way out of this difficulty. In the case just described, the HIV experts could imagine a hypothetical scenario in which others knew the information in question and then consider whether a majority in that hypothetical situation would vote to in favour of releasing that information in the actual world in which the information is not widely known. Thus, the experts would conduct a hypothetical vote on what should be done in the actual world rather than a hypothetical vote on what should be done in the hypothetical world.

[27] See, for example, Hooker (2002).

[28] A parallel can be drawn to one prominent justification for the authority of the law, due to Joseph Raz. That justification appeals to the same kind of consideration that we suggest could ground a norm against unilateral action: 'The normal and primary way to establish that a person should be acknowledged to have authority over another person involves showing that the alleged subject is likely better to comply with reasons which apply to him (other than the alleged authoritative directives) if he accepts the directives of the alleged authority as authoritatively binding, and tries to follow them, than if he tries to follow the reasons which apply to him directly' (Raz, 1994, p. 214; see also Raz, 1986: 38-69).

[29] As often with real world-examples, there are complications and qualifications. It might be objected that Galileo was not not faced with a true unilateralist situation since (i) it is unclear that there were numerous actors capable of undertaking the initiative that Galileo undertook, and (ii) one could question whether all the relevant parties were purely concerned with promoting the common good.

[30] This case might seem rather different to a unilateralist situation: one might assume that the mafia bosses are not motivated by the common good. However, they might take themselves to be so-motivated, and thus might take themselves to be in a unilateralist situation. This raises the possibility that they would attempt to comply with the principle of conformity, thus helping to sustain the cartel.