Social Epistemology Review and Reply Collective, 2016
Vol. 5, No. 4, 33-36. http://wp.me/p1Bfg0-2RY

Response to the Unilateralist's Curse
Stuart Armstrong, Oxford University

Due to the "Winner's curse", those that end up winning an item at auction may have paid too much for it. After all, in the estimation of all the other auction participants—which may include everyone in the entire world with any interest in that item—it wasn't worth that final price. The problem is that every bidder has a private, partial estimation as to the item's worth, and the person who wins is the one whose partial estimation is the most overly optimistic.

Once one is aware of it, there are ways to combat the winner's curse—most simply, by underbidding. But such "curses" are not limited to auctions. Consider geoengineering, for instance—the deliberate modification of the Earth's environment to combat climate change. Sulphate aerosols in the stratosphere are one such intervention. Like any similar ideas, it has advantages and costs.

It is, however, very cheap and easy for any country to pump enough sulphates into the stratosphere. Thus, if each country *independently* assessed the advantages and costs of the intervention, it would likely be started by the country who estimated the costs lowest and the advantages greatest. If we multiply the entities capable of starting a geoengineering project—most cities should be able to afford it, for instance—it becomes almost certain that someone will start it, even if most of the rest of the world thinks it's a terrible idea.

This is the "Unilateralist's curse" that Bostrom, Douglas, and Sandberg (2016) describe. In a situation where many actors could unilaterally initiate a project, they should reduce their estimations of the project's benefit, just as auction participants decrease their valuation of the item they're after (of course, if there are many actors any one of which could unilaterally *block* a project, they should conversely *increase* their estimation of benefit).

The paper's argument is easy to follow, and has the advantage of being true. They propose various norms and approaches that can be used to combat the curse, and bring individual action more in line with the true value of the project.

Beyond Unilateralism

My aim here is not to criticise the paper in theory, but to try and sketch its applicability in practice. The geoengineering example is not yet an issue, but it feels very unlikely that a single city will start a major project of global consequences, over the objections of the rest of the world. It feels unlikely, because we have experience of how these things go now: international agreements. Take pollution, for instance. This is more a question of coordination that unilateralism, but they follow the same pattern: long and tedious negotiations between various actors, until all powerful actors at least grudgingly accept the result, and the less powerful objectors are few and overruled.

In this case, international institutions and norms, created to solve coordination problems,

also serve to solve the unilateralist's curse: the process of negotiations would inevitably involve the sharing of information and benefit estimations.

There is a Burkean argument (named after the statesman and philosopher Edmund Burke) that the institutions and traditions of society reflect a collective wisdom about how things should best function and serve as a brake on dangerous tendencies that might otherwise develop. One doesn't need to take that philosophical position to observe that institutions clearly adapt to past problems, and take steps to prevent them from emerging again. And that this means that perfect individual rationality may not be needed.

Consider for instance a researcher that uses virtual screening (computer aided drug discovery) to find a promising anti-malaria molecule. But the same virtual screening flags the molecule as having possible side effects. Given that the researcher is in competition with many other virtual screeners, should they forward the hit to the next development stage?

Almost certainly they should. The process from promising molecule to full-blown drug is a long and arduous one, with many stages dedicated exclusively to ferreting out side effects and dangers. Stopping the process at that early stage makes little sense.

In a similar way, the functioning of a market society depends on innovators developing new ideas beyond the common wisdom. This means that innovators are encouraged to pursue projects that majority opinion thinks is worthless, and that are most likely worthless. The Unilateralist's curse may thus be a valid individual warning for individual entrepreneurs ("you are unlikely to succeed") but fails as an overall judgement of the entrepreneurship ("it is good that your try").

We can multiply the examples, but they all fit in a broad pattern. The Unilateralist's curse is dedicated to figuring out the best epistemology for situations—how can we figure out the best action according to the best possible use of our partial information? But society is not dependent on everyone following best practices or having the best possible knowledge (rather the opposite, in many cases); instead it often designed to make use of biases and errors to achieve desirable ends (think of adversarial systems, such as the legal system, politics, the market, academia...). Therefore it is not surprising that biases such as the Unilateralist's curse turn out to have manageable and managed impacts in many situations.

Thus the real question is rather: in what situations is it important to take the Unilateralist's curse into account? There are three clear examples: tail risks, under-regulated areas, and new technologies.

Tail Risks

Tail risk concerns the small probability of extremely detrimental impact. Since such disasters happen so rarely, and memories fade in time, society as a whole is generally extremely ill-prepared for such extreme events. So, for instance, consider the recent gain-

of-function research for flu viruses. The tail risk is that these viruses escaped containment (an unfortunately not uncommon event, even in high security labs) or the experimental protocols were used by bioterrorists or governmental bio warfare divisions. When challenged, the researchers tended to rely on the standard tools of science safety assessments: internal reviews, reviews by researchers with closely aligned priorities, and general arguments about the importance of the pursuit of knowledge. They also, quite naturally, fought to defend their own research corner—a natural, and indeed obligatory action in today's competitive research arena.

These tools, however, were constructed to deal with standard research problems. There is no reason to suspect that they were finely calibrated to cope with large tail risks. Standing up for one's project and the general pursuit of knowledge are laudable positions for most researchers to take, but gain of function research clearly is in a different category, where more care needs to be taken. Tail-risk is also an area that by definition will be hard to model: little data to base it on, and hence it becomes very model-driven in its conclusions. That in turn means that different actors will get very different views on what should be done, and the curse bites harder.

Under-Regulated Areas

Let's turn now to under-regulated areas. Lead in gasoline is an example of an innovation with potential security hazards that turned out to have disastrous impacts. The governments of the 1920s were not setup to catch and regulate pollution with any degree of efficiency, so it was clear that any detrimental effect would go on unchallenged. The designer of the product had nothing to catch them out if they produced anything dangerous; therefore they should have taken much more care with their designs.

Nowadays, governments are more likely to regulate and control products, so it's more a question of whether such regulations are likely to be effective. If there's a known flaw in the regulatory system—through corruption, mismanagement, or regulatory capture—then the companies and designers need to take that responsibility on themselves.

New Technologies

But sometimes the issue is not flawed regulation, but no regulation. Which brings us to the issue of new technologies, which are seldom pre-emptively regulated (and if they are, are rarely pre-emptively *well* regulated). The risk is low if any problem can be easily and rapidly regulated away or controlled by the manufacturer. In other cases, the risk is potentially high, and may connect with tail risk for certain innovations, such as biological or computerised ones.

A large tech company like Google, FaceBook, Microsoft, or Apple, has the potential to roll out an innovation to a large amount of people at once. If this innovation is very popular yet detrimental to a large class of people (one can think of some way of automating an entire profession, possibly) then it is unlikely to be successfully regulated away. Therefore it's important that the tech company assess the costs and benefits of their

innovation, using considerations like the Unilateralist's curse (which doesn't mean they shouldn't do it, just that they shouldn't rely on their own naïve initial judgement). Smaller companies, whose products are more likely to be retrospectively regulated away, have less need to pay attention to this risk.

Thus the strength of the Unilateralist's curse applies the most in areas where society has not pre-installed protective measures. In these areas (such as future—not current—artificial intelligence, which is a simultaneous under-regulated tail risk new technology) extra caution is necessary, and using the tools of formal epistemology, including awareness of the Unilateralist's curse, become necessary. The less useful past experience is to the problem, the more vital our theories become.

Contact details: stuart.armstrong@philosophy.ox.ac.uk

References

Bostrom, Nick, Thomas Douglas, Anders Sandberg. "The Unilateralist's Curse and the Case for a Principle of Conformity." *Social Epistemology* (2016): 1-22. doi: 10.1080/02691728.2015.1108373.