Prediction Markets for Science: Is the cure worse than the disease? (Draft)

Mike Thicke

March 15, 2017

1 Introduction

In a prediction market, buyers and sellers exchange contracts whose value depends upon the fulfillment of a prediction. The most common prediction markets are political, with contracts based upon the results of democratic elections. For example, a contract might be worth \$1 if Hillary Clinton wins the 2016 U.S. presidential election and \$0 otherwise. Prices in such markets can be interpreted as predictions about the likelihood of a particular event. If contracts priced at \$0.70 pay out \$1 70% of the time, then the expected value of each is \$0.70, its price. Conversely, a price of \$0.70 can be interpreted as implying a 70% probability that the associated event will occur. Favour.

The Iowa Electronic Markets,[1] operated by the University of Iowa, has been running election prediction markets since 1988, and its predictions have proven remarkably accurate, outperforming both expert predictions and opinion polls. Not long after the IEM's founding, economist Robin Hanson proposed using prediction markets for science, and several others have made similar proposals (Hanson 1995; Bell 2006; Potthoff 2007; Pfeiffer and Almenberg 2010; Hsu 2011). Recently, prediction markets featured prominently in the "Reproducibility Project," which sought to replicate the results of psychology experiments. The markets, in this case, successfully predicted the outcome of 71% of replication attempts (Dreber et al. 2015).

Prediction markets offer to circumvent genuine problems with scientific research, particularly with peer review and measures of scientific consensus. Peer review has been subjected to a number of criticisms, from unreliability to the exclusion of minority viewpoints. Consensus alone does not indicate scientific knowledge, as consensus can arise for any number of reasons beyond the weight of evidence, and current methods of demarcating epistemically justified from unjustified consensus are insufficient. Methods of amalgamating evidence directly, such as through meta-analysis, have similar problems. Science prediction markets, if they worked as advertised, could help to circumvent these problems. However, there are strong reasons to believe they would not work as advertised.

[1] http://tippie.uiowa.edu/iem/

Although proposals for incorporating prediction markets into scientific research typically include some discussion of possible or perceived challenges, there has been no serious examination of their potential problems for scientific practice. Prediction markets for science are almost exclusively discussed by their proponents. The primary aim of this paper is to correct for this imbalance by offering three critiques of prediction markets for science. First, it will argue that prediction markets for science could be uninformative or deceptive because scientific predictions are often long-term, while prediction markets perform best for short-term questions. Second, it will argue that prediction markets could produce misleading predictions due to their requirement for determinable predictions. Prediction markets require questions to be operationalized in ways that can subtly distort their meaning and produce misleading results. Third, it will argue that prediction markets offering significant profit opportunities could damage existing scientific institutions and funding methods.

Throughout the paper I will appeal to climate science as a potential test case for science prediction markets. This is both because climate science is of great public interest and because climate science is inherently predictive; a primary objective of climate scientists, as a group, is to predict how the climate will change in the future. The focus on quantitative measurement and prediction in climate science makes it a near-ideal test case for prediction markets. Climate science is also decentralized and multidisciplinary, so the aggregating potential of markets should be more important for it than for other large-scale endeavours, such as particle physics, that are more centralized.

2 Prediction markets address a genuine need

Advocates for employing scientific prediction markets argue that they have the potential to solve significant problems in the current operation of scientific research. Robin Hanson, for instance, levels a series of charges at the current practice of science (Hanson 1995). "Peer review," he claims, "is just another popularity contest" (Hanson 1995, p. 4). It takes too long to expose error. Publication quantity is often valued over quality. Grants reflect "insiders"' opinions of past research rather than the correctness or value of future results. In debates relevant to public policy, such as over global warming, "an honest consensus of relevant experts is often lost from public view, as advocates on each side accuse the other of bias and self-interest" (Hanson 1995, p. 5). Shi-Ling Hsu argues that prediction markets for climate science could free it from "ideological bias" (Hsu 2011, p. 106). Tom Bell argues that the "mass media" often oversimplifies and sensationalizes scientific reports (Bell 2002). Dreber et al. (2015) argue that prediction markets could solve problems of publication bias and replication failure. A serious discussion of prediction markets for science must acknowledge that these complaints about science and its role in public debate, while perhaps exaggerating the problems with current scientific practice, are not spurious.

This section will consider two pillars of scientific authority: peer review and consensus. These are the most-cited indicators of credible scientific opinion, and

consequently the most frequently criticized.

According to Ziman (2004), peer review is the

> principal social mechanism for quality control in academic science... [It] keeps the official scientific literature reasonably honest and factually reliable. It favours precise, thorough and cogent argumentation and sets high benchmarks for technical performance. (Ziman 2004, p. 42)

Peer review, in Ziman's account, is what ensures the quality and trustworthiness of scientific publications. If a scientific claim is not based upon peer reviewed research, then it ought to carry little weight in public debate. In contrast, peer reviewed research, while not infallible, carries the stamp of scientific authority.

While peer review is meant to ensure the quality of individual contributions, consensus measures report aggregate scientific opinion. An overwhelming consensus, such as the often-reported 97% consensus on anthropogenic global warming, is meant to grant credibility to the underlying claim: if 97% of climate scientists believe in human-caused warming, there is a very high chance that it is indeed occurring (Cook, Nuccitelli, et al. 2013). In a meta-analysis of consensus estimates in climate science, which finds that there is near-universal consensus on global warming, Cook, Naomi Oreskes, et al. (2016) concludes: "The level of scientific agreement on [anthropogenic global warming] is overwhelmingly high because the supporting evidence is overwhelmingly strong" (Cook, Naomi Oreskes, et al. 2016, p. 6). Consensus stands-in for evidence, which in turn justifies the claim at hand. Further, peer review and consensus work together to establish credible scientific opinion; assessments of consensus such as Oreskes (2004) are often based on surveys of peer reviewed literature. Finally, as publishing in peer reviewed journals is obligatory for scientific career advancement, consensus measures of academic scientists are necessarily measures of the views of those who produce peer reviewed research.

The first pillar of scientific authority, peer review, originated in the seventeenth century as a form of state-sponsored censorship (Biagioli 2002). The Royal Society in Britain and the Academie Royale des Sciences in France instituted peer review to ensure that articles published in their journals would not contain controversial material that would damage the societies' social position. Peer review was thus originally intended not to protect readers from low quality research, but to protect the academies. Although peer review has ceased to be directly tied to protecting political interest, many argue that it retains a conservative bias. Horrobin (1990), for instance, documents many instances in which research that was eventually judged to be seminal, and sometimes even Nobel Prize winning, was initially rejected by reviewers. Horrobin argues that, "we must take seriously the possibility that we have traded innovation for quality control" (Horrobin 1990, p. 1439). Nicholson and Ioannidis (2012) found that National Institute of Health (NIH) insiders are very likely to receive NIH funding despite having, on average, "modest" citation records, while a majority of authors of highly cited research do not, possibly suggesting that "study-section

members fund work that is more similar to their own, or that they are chosen to serve as study-section members because of similarities between their own and funded grants" (Nicholson and Ioannidis 2012, p. 35). "A Social Epistemological Inquiry into Biases in Journal Peer Review" (n.d.) argues that researchers seeking to publish in high prestige journals, with correspondingly high rejection rates (especially early career researchers who depend upon relatively rapid publication), might "adjust their manuscripts to match the assumed preferences of reviewers" ("A Social Epistemological Inquiry into Biases in Journal Peer Review" n.d., p. 22). If those reviewers, as is often the case, are established researchers, then this would discourage scientists from submitting research supporting controversial or unpopular views.

Compounding conservative pressures on early career researchers, reputational bias in peer review favours submissions from established researchers. Peters and Ceci (1982) submitted articles to several top psychology journals that had previously been accepted by the very same journals, with only the names and institutional affiliations of the authors altered. Only a small proportion of editors or reviewers detected the duplication, while eight of nine undetected articles were rejected, with the most common reason being faulty study design or statistical analysis (Peters and Ceci 1982, p. 190). Peters and Cecil speculate that "when referees examine a manuscript submitted by researchers working at highly respected institutions, they may be more sensitive to making 'false negative' evaluations, that is, rejecting papers of quality, whereas the major concern in reviewing papers of individuals from lesser known institutions may be that of avoiding 'false positive' errors, that is, accepting flawed work" (Peters and Ceci 1982, p. 192).

Fitzpatrick (2010) argues that conservative biases in peer review are unsurprising because "those faculty and administrators who are in the position of performing assessments of the careers of other, usually younger, faculty are of necessity those who have sufficiently benefitted from the current credentialing system as to rise to that position" (Fitzpatrick 2010, p. 171). This explanation evokes Merton and Zuckerman's (1971) picture of peer review as enacting "a hierarchic structure in which power and authority are largely vested in those who have acquired rank through cumulative scientific accomplishment" (Zuckerman and Merton 1971, p. 81). In this account, peer review entrenches accepted scientific theories and methods by subjecting those that violate orthodoxy to disproportionate scrutiny.

Beyond enacting a conservative bias, peer review has been charged with unreliability. Rothwell and Martyn (2000), studying the results of peer review in clinical neuroscience, conclude that "in neither of the journals that we studied was agreement between independent reviewers on whether manuscripts should be published, or their priority for publication, convincingly greater than that which would have been expected by chance alone" (Rothwell and Martyn 2000, p. 1966). Richard Smith, former editor of the *British Medical Journal*, charges that "in addition to being poor at detecting gross defects and almost useless for detecting fraud, [peer review] is slow, expensive, profligate of academic time, highly subjective, something of a lottery, prone to bias, and easily

abused" (Smith 2006). While Lee et al. (2012) rightly observe that many accusations of bias or unreliability in peer review appeal to an unrealistic concept of "true quality" against which biased reviews are judged, it is hard to see peer review as a credible indicator of trustworthy or authoritative research if the agreement between reviewers is little better than what would be "expected by chance alone."

Consensus is the second pillar of scientific authority, but the inference from consensus to evidence to truth is not as direct as invocations such as Cook, Naomi Oreskes, et al. (2016) suggest. While consensus measures refute claims that there is significant disagreement about global warming among climate scientists, it would be a mistake to infer from consensus alone that human-caused global warming is indeed happening. That 97% of climate scientists agree about global warming does not imply that there is a 97% chance that human-caused global warming is occurring. It is unknown whether 97% of scientists believe that there is a 60% chance of this being true or that there is a 99% chance of it being true. And there is no way of judging, based upon just this metric, whether those scientists are correct to believe what they do. Examples of mistaken consensus abound in science. For instance, scientists were mistaken that species are immutable, that the Earth sails through an ether sea, and that the continents are permanently fixed in their positions on the Earth's surface. There is simply no way to directly infer a probability estimate regarding the truth of a scientific theory from the proportion of scientists who endorse it. Nevertheless, the 97% consensus measure has significant epistemic and rhetorical appeal. It is a simple quantity that can be easily understood without any climate science expertise; anyone can see from it that there is overwhelming agreement about global warming among climate scientists. Consensus measures are not useless; they are insufficient. What is needed is a method of distinguishing consensuses that indicates scientific knowledge from consensuses that do not, and ideally a method that indicates what degree of confidence in the associated theory is justified by a scientific consensus. This is what prediction markets offer to achieve, while all direct evaluations of consensus fail to do so.

Consensus measures are useful to the extent that they allow for a cognitive division of labour. If they can be trusted as reliable indicators of the truth, they allow non-experts to adopt informed beliefs without having detailed knowledge of the underlying subject matter. This is a very important function–it is impossible for *anyone*, let alone non-scientists, to gain the expertise necessary to independently evaluate the evidence in more than a few scientific disciplines. Given that democracy requires voters to have informed opinions about an increasing variety of scientific claims, some method of evaluating those claims without detailed expertise is essential. Methods of distinguishing between epistemically justified and unjustified consensus ought to maintain this division of labour. If assessing a consensus requires examining the underlying scientific arguments and evidence, the consensus itself loses epistemic significance; one might as well just examine the evidence instead of assessing the consensus. Assessments of consensus also ought to be practically applicable; if the requirements are so stringent as to be unrealizable for any real-world consensus, or too vague to be

applicable in actual circumstances, they are of little use. Finally, methods of assessment ought to be reliable: they should admit consensuses that do reflect knowledge while rejecting those that do not, and the degree of confidence they attribute to a consensus should reflect its actual reliability. Several methods of assessing consensus have been proposed, but none fully satisfy these criteria—division of cognitive labour, practical applicability, and reliability. I will discuss three representative attempts.

Longino (1994) identifies four attributes of a community that can be expected to reach an epistemically justified consensus (Longino 1994, pp. 144–45). First, there must be recognized avenues for criticism. Second, the community holding the consensus must be responsive to criticism. Third, the community must share standards of evidence. Fourth, the community must be intellectually egalitarian; consensus cannot be the result of political or economic power or other forms of coercion or exclusion. These criteria do accomplish a cognitive division of labour: no assessment of the underlying scientific theories and evidence is required to assess the consensus. However, it is hard to see how they could be practically applied. Longino recognizes that these criteria are only achievable by an "idealized epistemic community" (Longino 1994), but claims that achieving them is a matter of degree. Presumably the more a community realizes these criteria the more its consensus claims ought to be trusted. But to what extent do these criteria admit of degree? How responsive is the climate scientist community to dissent? To what extent do climate scientists share standards of evidence? Climate science is not perfectly egalitarian; university researchers publishing in peer-reviewed journals have much more intellectual authority than their unaccredited critics. Longino's criteria seem to be good criteria for an epistemic community to aspire to, all things being equal, but are not particularly useful for evaluating the consensus beliefs of actual communities. Biddle (2007) observes, for instance, that Longino's criteria are too abstract to diagnose concrete problems with privately-funded clinical trials. Conversely, Pinto (2014) argues that Longion's criteria could be used to discount the scientific consensus on global warming, as scientists cannot in practice respond to all criticism from all sources, and cannot be completely egalitarian. The debate over whether global warming is occurring is closed, despite vigorous objections from outside mainstream academia.

Tucker (2003) proposes criteria that are more practical to apply than Longino's, but are vulnerable to charges of unreliability. He requires that a consensus-bearing community be "uncoerced", "uniquely heterogeneous" and "sufficiently large" (Tucker 2003, p. 504). Tucker proposes that we should infer knowledge from consensus when there is no better explanation for that consensus. If the consensus appears to be the result of political coercion, as was the case for the Soviet consensus around Lysenkoism, we should not infer knowledge from that consensus. The "uniquely heterogeneous" condition requires that the community share no common property that could explain its common belief independently of possessing shared knowledge. For example, Solomon (2001) argues that Barbara McClintock's evidence for non-Mendelian inheritance in corn was largely ignored by other geneticists partly because of sexism; she reports one

male geneticist describing McClintock as "just an old bag who'd been hanging around Cold Spring Harbor for years" (Solomon 2001, p. 113). The scientific community was homogeneously male, which better explains their maintaining a Mendelian consensus in the face of McClintock's contrary evidence than shared knowledge about some potential flaw in McClintock's study. Tucker's method is more practically applicable method than Longino's primarily because it is comparative; it uses his three criteria to evaluate competing explanations, and if the best explanation is knowledge, we should trust the consensus. This is certainly an improvement, but it is hardly deterministic, as whether a consensus is sufficiently large or uncoerced will often be a matter of opinion.

Further, building upon Solomon (2001), Miller (2012) argues that Tucker's criteria could admit consensuses that are not due to shared knowledge and not due to a single biasing factor, but rather due to a diverse set of biases that conspire to lead scientists to the same conclusions. Tucker disqualifies consensuses in which the community shares a common biasing factor, but argues that it is unlikely, though possible, that multiple independent factors could lead scientists to the same conclusion. Miller disagrees, arguing that "different groups often have mutual interests which will cause them to reach a consensus on occasion on a particular matter despite disagreeing on other things" (Miller 2012, p. 1308). The academic community has a documented liberal bias, with one survey finding that 60% of academics at American higher education institutions identify as far-left or liberal, while only 13% identify as conservative or far-right (HERI 2014, p. 39). A survey of climate scientists found similar numbers: 67.5% of surveyed scientists identified as liberal while 13% identified as conservative (Rosenberg et al. 2009, p. 314). Perhaps this liberal bias, combined with fear amongst conservative-leaning climate scientists that voicing dissent will harm their career prospects, explains the global warming consensus better than shared knowledge, despite there being no single homogenous factor. If it is common for biases to combine to generate consensuses that are not knowledge-based, then Tucker's method will not be reliable. Miller argues that "apparent consilience of evidence" should substitute for Tucker's unique heterogeneity condition (Miller 2012, p. 1309). Miller's condition requires that multiple lines of evidence all point to the same conclusion. This condition is similar in spirit to Solomon's claim that consensus is justified only when one theory is supported by all available empirical evidence (Solomon 2001, p. 119). According to Miller, Tucker's unique heterogeneity is a proxy for consilience of evidence; different social groups tend to emphasize different sorts of evidence, and so ensuring that the community is not dominated by any single social group ensures that it is not dominated by any single kind of evidence. Given the problems with unique heterogeneity, therefore, it makes sense to appeal directly to apparent consilience. This may well be right, but it comes at the cost of violating the cognitive division of labour. Assessing consensus according to Miller's or Solomon's methods require a detailed assessment of the evidence at hand, which will likely require significant domain expertise. This means that only climate scientists, or those with a significant knowledge of climate science, will be qualified to judge whether the global warming consensus indicates shared knowledge. This

significantly diminishes the appeal of Miller's approach.

An alternative approach to consensus measures for assessing scientific opinion is to examine the evidence directly. A common way to do this is through a meta-analysis, in which the evidence from a number of studies is amalgamated to produce a quantitative result. Meta-analyses are often used in medical research, for example, to assess the effectiveness of a treatment method. Meta-analysis shares with prediction markets and consensus measures the advantage of making an easily-understood and communicated assessment of a hypothesis. However, it fails according to both the division of labor and reliability criteria. Meta-analysis requires significant domain expertise to perform, and so cannot be effectively performed by outsiders. This means that trusting a meta-analysis requires trusting the expert who performed it–trusting that she considered all of the available evidence, that her method of aggregating the data is impartial, and so on. If a universal meta-analysis algorithm existed that did not require domain expertise and was generally accepted as reliable and unbiased, then meta-analysis would be a strong alternative to prediction markets. However, as Stegenga (2011) argues, such a universal method is impossible for meta-analysis or any other quantitative method of amalgamating evidence. Douglas (2012) consequently proposes a qualitative, explanatory approach to evaluating evidence. In her approach, scientists holding competing views are asked to explain the available evidence, and their explanations are assessed according to a number of criteria including completeness, internal consistency, and predictive potential (Douglas 2012, p. 152). The reliability of Douglas's method will depend upon the abilities of the assessor(s) to competently and impartially assess competing explanations, as well as their ability to consider all relevant explanations. If only explanations from peer reviewed literature were considered, for example, then the biases of peer review would necessarily be reflected in the assessment's outcome.

A cognitive division of labor for these methods, as well as for Miller's and Solomon's, could perhaps be achieved in a two-step process: first, domain experts assess the consensus or evidence, and second, non-experts assess the initial assessment process. Whether this division of labor can be achieved will depend upon whether the non-experts are able to competently judge the expert assessment without themselves being capable of making it. Miller might look to whether the assessment appealed to consilience, while Douglas might check whether it required internal consistency for each of the competing explanations. Such an approach could be feasible, but I am skeptical whether non-experts could reliably distinguish between cases where expert assessors *really did* employ the desired method and cases where they *merely appeared* to do so. In cases where there is no particular reason to suspect such deceptive appearances to be at play this might not be a significant worry, but in ideologically charged cases such as with global warming, it certainly would be.

None of these methods is practically applicable and reliable, while maintaining a cognitive division of labour. This does not make them meritless; they all offer ways of discussing the epistemic merits of consensuses. But they are not satisfactory for evaluating particular cases of consensus, such as the consen-

sus of anthropogenic global warming by non-scientists.[2] Prediction markets, if they work as advertised, satisfy all of these conditions. They can be applied to practical questions, and the interpretation of prices as probabilities requires no domain expertise. They have proven very reliable in some domains, and there are some indications that they could be similarly reliable in science. Prediction markets that are open to everyone can avoid the exclusionary aspects of peer review, while obviating the need for (potentially biased) quality control.

3 The promise of prediction markets

Prediction markets have the potential to improve upon the current system of assessing scientific questions through consensus measures of peer reviewed literature and surveys of accredited experts. Prediction markets provide simple probability estimates that require no domain expertise to understand and have proven reliable in some domains. This section will lay out the case for employing prediction markets to assess scientific questions.

The best known and most studied prediction markets are political markets, and especially the Iowa Electronic Markets (IEM) run by researchers at the University of Iowa College of Business. Such markets have proven to be remarkably accurate predictors of election results (Forsythe et al. 1992; Wolfers and Zitzewitz 2004; Sunstein 2006). Although prediction markets outside of the political sphere are still relatively rare and small-scale, there is some evidence that they perform well in other realms, even when conducted for imaginary, rather than real, money (Pennock et al. 2001).

Prediction markets for science have a long history, though all have been small-scale and most have been for imaginary money. The Foresight Exchange[3], for example, has been operating since 1994, with imaginary money markets covering a wide range of scientific topics including the level of global warming by 2030, the ambient level of CO_2 in the atmosphere by 2030, and whether the cosmological constant is greater than 0. New Zealand-based real-money prediction market iPredict has contracts addressing whether a new element will be added to the periodic table by the end of 2017 and whether any extinct New Zealand bird will be rediscovered by the end of 2017. None of these markets has many participants (the Foresight Exchange global warming market had only 7 trades in 2015, for example).[4]

Some limited experiments on science prediction markets have been performed. Almenberg, Kittlitz, and Pfeiffer (2009) created an experimental prediction market where participants invested in six mutually exclusive hypotheses

[2]Miller's criteria have been cited by climate scientists as supporting the anthropogenic global warming hypothesis (Jacobs 2014). This demonstrates the practical applicability of Miller's criteria, but not their amenability to a cognitive division of labor.

[3]http://www.foresightexchange.com/

[4]Pennock et al. (2001) found a strong correlation between Foresight Exchange prices and observed outcome frequencies, though it is unclear how many of the markets they examined were science related.

about a fictitious biochemical pathway. Participants were presented with information over the course of the experiment and in different ways (information could be public, private, or private-then-public) and the experimenters tracked the reaction of market prices to this information. When information was public or private-then-public, mispricing relative to an idealized Bayesian reasoner was low, but when information was kept private, there was significant mispricing (Almenberg, Kittlitz, and Pfeiffer 2009, p. 5). Thus the accuracy of science prediction markets could depend in part on the extent to which scientists share their data through publication, and in consequence could depend upon the reliability of peer review. Dreber et al. (2015) ran prediction markets to predict the results of replication attempts of psychological studies, as part of the highly-publicized Reproducibility Project: Psychology.[5] Prediction markets predicted the outcomes of replication attempts 71% of the time while simple surveys of the market participants made correct predictions only 58% of the time (Dreber et al. 2015, pp. 15344–5). This suggests that prediction markets could outperform surveys even if the markets are limited to accredited participants.

Proposals for science prediction markets are much more ambitious than these limited experiments. Robin Hanson proposes an "idea futures" market, where banks issue mutually-exclusive contract pairs for a price, say $10 per contract, that will then be judged by an impartial party at a specified date (Hanson 1995). For instance, a contract pair might be "it will rain tomorrow" and "it will not rain tomorrow." Each of these contracts would pay $10 (minus a fee) if the prediction came true. Since only one prediction could come true, the bank would take no risk. Once an individual has purchased a contract pair, he or she could sell one side while keeping the other, anticipating that the price of that side will rise in the future (since the contracts are mutually exclusive, their market value should vary inversely). More scientifically interesting contracts could include surface temperature measurements at specified dates, frequency of extreme weather events, and so on. In this way, prediction markets for science could be implemented using existing institutions, though U.S. law would have to change to accommodate such contracts (Arrow et al. 2008; Bell 2002).

Hanson envisions a scenario where Alfred Wegener could have employed idea futures when faced with widespread dismissal of his hypothesis that continents drifted on the Earth's surface. Hanson claims that Wegener's 1915 theory was ignored at least partially because it would have been a poor career choice for others to support his claim, even if they found his evidence compelling (Hanson 1995, p. 9). If Hanson's markets existed, Wegener could have taken out contracts predicting that his theory would eventually be confirmed. Presumably he could have gotten a very good price from other scientists, as none judged his theory credible. But as Wegener created more and more contracts, others who privately

[5]The Reproducibility Project: Psychology is organized by the Center for Open Science (https://cos.io/). Between 2011 and 2014, over 270 contributors attempted to replicate 100 published psychology studies. They successfully replicated 36% of those studies (Open Science Collaboration 2015). This result has been widely publicized as an indictment of experimental psychology and scientific methodology in general (Carey 2015). Whether this is a fair interpretation of their results is still a matter of debate (Gilbert et al. 2016).

judged his claims credible could also buy contracts, and perhaps even pursue further research in hopes of realizing a profit. If they did so anonymously, they could have done so without fear of academic derision. In this way, the price of continental drift contracts might have revealed a probability estimate different from what the anti-drift consensus suggested and might also have hastened the discovery of new evidence as scientists sought to profit from their investments. For Hanson, this exemplifies the promise of prediction markets.

Shi-Ling Hsu proposes a more complicated market than Hanson's, specifically dealing with climate change predictions (Hsu 2011). Hsu's proposal consists of two parts. First, he proposes a variable carbon tax indexed to an aggregation of climate indicators (temperature, sea level, frequency and severity of climate events...). Second, he proposes a market for emissions permits that would grant holders exemptions from the climate tax. These emissions permits would be for specific years in the future, and so their price should vary according to expectations of the carbon tax, which in turn should depend on expectations of the climate variables used to determine the carbon tax. Therefore, the emissions permit market could be considered a prediction market for climate outcomes.

In addition to Hanson's and Hsu's proposals, some have proposed that current funders of science such as the National Science Foundation (NSF) could supplement or replace their current funding procedures with subsidized prediction markets (Almenberg, Kittlitz, and Pfeiffer 2009). For example, contract pairs might be sold for $1 but pay a total of $2 upon expiration. Or there might be some non-linear mapping from prediction market success to real-money payoff. Markets like this would have to be limited to a certain group of participants to avoid simply funding recreational gamblers or speculators, but even such limited markets could be useful for yielding useful predictions and encouraging research. They could also help to fund research, as scientists who trade on the market based on information discovered during their research should be able to profit, and those profits could be amplified by funding agencies. In this way, scientists with the most significant research findings would reap most of the rewards from trading and would be best able to produce further research.

Hsu argues that climate science is particularly suited for prediction markets because of its broad, decentralized nature. He goes so far as to claim,

> There is no better mechanism for processing climate science than prediction markets, and there is no better way to showcase the power of prediction markets than to apply one to climate science. (Hsu 2011, p. 106)

There are good reasons to use climate science as a test case for prediction markets and for prediction markets to be used for climate research. Understanding how prediction markets could benefit climate science should also help to understand how they could benefit areas of science.

Prediction markets require testable predictions. Although Hanson suggests that prediction market judges may assign non-binary evaluations of predictions,

this seems fraught with problems. (A non-binary judgment could, for example, grant $0.80 to "yes" contracts and $0.20 to "no" contracts rather than $1 and $0 respectively in the case of an ambiguous result.) It is difficult to see how such judgments could be made immune from charges of ideological bias or conflict of interest, as they would rely on the judgment of a single individual. Better are predictions that are amenable to transparent and unambiguous empirical test.[6] While measuring the climate is never completely straightforward (Edwards 2010), variables such as surface temperature and precipitation should be measurable without undue controversy. The apparent prevalence of salient measured quantities in climate science seem to make it an ideal candidate for science prediction markets. If climate science prediction markets fail to perform, prediction markets for scientific disciplines without as much emphasis on quantitative measurement will likely face even more difficulty.

Hsu highlights two reasons that prediction markets should be beneficial to climate science. First, he argues that climate science research is decentralized and markets are at their best when aggregating decentralized knowledge. Second, he argues that prediction markets incentivize individuals to reveal information free of ideological bias, and current disputes over climate science seem especially prone to such bias.

Climate science is of course not the only large, interdisciplinary field of science. The physicists and engineers who run massive physics experiments such as the Large Hadron Collider at CERN must coordinate the efforts of thousands of people who may know very little of each other's work (Galison 1997; Knorr-Cetina 1999). However, those scientists usually have a central organization, such as CERN, coordinating their activities. While the International Panel on Climate Change's periodic assessments play a significant role in setting the climate research agenda, it has no direct role in funding or directing such research. Funding climate research through prediction markets could both help solve climate science's data aggregation challenges and encourage scientists to cooperate without the need for a central funding and planning agency.

Hsu argues that prediction markets could also encourage scientists to publish results free from "ideological bias." As discussed above, it is well established that there is consensus among climate scientists that anthropogenic global warming is occurring. However, one cannot argue simply and straightforwardly from consensus to knowledge. If there is some other explanation for that consensus, say that all climate scientists share a liberal, pro-regulation ideology, or that a cabal of elite climate scientists suppresses dissent, there is reason to question the relevance of that consensus for arriving at conclusions about the climate. Even if one believes that charges of pervasive ideological bias against climate science are delusional, one might support prediction markets because they could fortify climate science against charges of bias. There are two apparent ways that prediction markets might help to do this.

First, election prediction markets have proven robust to political biases.

[6] I am not aware of any real-money prediction markets that implement Hanson's non-binary system.

Communities of scientists are likely biased in a variety of ways, including politically. This would be no less true for prediction market participants, though perhaps such markets would attract participants from a wider range of political viewpoints. Existing prediction markets have similar biases: a majority of participants in the 1988 IEM presidential election market were more supportive of the Republican candidate, George Bush, than his Democratic opponent, Michael Dukakis (Forsythe et al. 1992). This was reflected in their trading behaviours. Bush supporters tended to interpret events, such as presidential debates, as favourable to Bush's chances, while Dukakis supporters interpreted the same events as favourable to his chances. Since there were more Bush supporters than Dukakis supporters, average opinion overestimated Bush's chances of victory. However, the actual market predictions were remarkably accurate, only undervaluing Dukakis by about 0.4%. Forsythe et al. (1992) argue that this was most likely due to a small group of "marginal traders" who correctly interpreted events and profited off of their peers' biases, while simultaneously bringing market prices into line with reality. Regardless of the mechanism, election prediction markets have repeatedly been shown to produce very accurate predictions, and so seem able to overcome the aggregate biases of their participants. Therefore even if there is an ideological bias in climate science, prediction markets could overcome that bias. If prediction market prices supported the current scientific consensus, this could foreclose arguments that the scientific consensus merely reflects the biases of climate scientists.

Second, prediction markets could circumvent other scientific institutions that have the potential to bias scientific conclusions, such as peer review. Few would agree that peer review is "just another popularity contest," but there are justified worries about the effect of peer review on the reliability of scientific publications. If scientific prediction markets are open to any participant, systemic biases among particular groups of scientists would merely represent profit opportunities for more rational traders, and market prices would be a better indicator of the truth than surveys of peer reviewed publications. Again, if market prices agreed with scientific publications, this would be strong evidence that the peer review process is not as subject to bias as its critics allege.

The potential to overcome biased consensus and avoid the potential problems of peer review makes a compelling case for adopting science prediction markets. Climate science is an ideal test case for these markets, though other areas of science with a significant emphasis on testable predictions could also be amenable to prediction markets.

4 Problems with science prediction markets

Prediction markets are markets, operating much like the stock market or any other financial market. They are thus susceptible to many of the same problems as any other market, such as bubbles, crashes, and manipulation. Sunstein (2006) describes one example of a prediction market bubble:

> In 2005, it was widely rumoured that Chief Justice William Rehn-

> quist would retire shortly after the end of the Supreme Court's term in June. An informational cascade quickly arose. People said that the chief justice would resign, not because they knew, but because other people said that the chief justice would resign. The cascade reached influential members of the media and even the U.S. Senate, leading them to join and hence to amplify the cascade. Prediction markets similarly foresaw his retirement. Nonetheless, he elected to stay on the bench until his death in September. (Sunstein 2006, p. L1962)

If we combined all of the evidence possessed by traders on Sunstein's prediction markets, it would amount to no more than idle speculation. Market participants were trading on rumour, and in doing so they drove up prices without epistemic justification, just as speculation over the profit potential of internet companies drove the NASDAQ to unheard of heights before an abrupt crash in mid–2000, in what became known as the "dot-com bubble." The Rehnquist market crashed when the judge did not, in fact, retire.

A notable instance of manipulation occurred during the 2012 US presidential election, when a trader on the popular Intrade prediction market succeeded in holding Obama at an implied 70% chance of victory even as other markets and polls suggested a much greater probability of Obama winning (D. M. Rothschild and Sethi 2015, p. 21). This manipulation, though, cost the trader nearly seven million dollars, and was limited to a single market. This suggests that while possible, manipulation of election prediction markets is quite difficult. However, for science prediction markets where the date of reckoning might not be so near, manipulation could be more effective.

Although prediction markets have generally performed favourably compared to expert predictions and polls, they have not done so universally. For instance, statistician Nate Silver's FiveThirtyEight prediction model, which incorporates a wide range of statistical data including polls and economic indicators, slightly outperformed prediction markets in the 2008 presidential election (D. Rothschild 2009, p. 897). Prediction markets performed worse than polling in predicting the 2016 Republican presidential nomination and the results of the "Brexit" referendum (*Who said Brexit was a surprise?* 2016).

The problems discussed so far are not particular to science prediction markets; they are challenges that all markets face. None of them is acute; predictions markets need not be epistemological panaceas or miraculous seers to be useful. Despite their problems, prediction markets generally perform well. However, prediction markets have only been implemented on a large scale for elections and sporting events, both of which have uncontroversial resolutions and known, short time horizons. Neither of these qualities applies generally to science prediction markets, and so inferring from the success of election and sporting prediction markets to expected success for science is unjustified. For the remainder of this section I will discuss problems that apply especially to science prediction markets. I will discuss three problems: long-term 50% bias, untenable operationalism, and outsized institutional consequences.

Even with rational traders who correctly assess the relevant probabilities, binary prediction markets can be expected to have a bias towards 50% predictions that is proportional to their duration. Prediction markets with more distant expiration dates, such as the Foresight Exchange global warming market that resolves in 2030, will have prices closer to $0.50 (when a correct prediction pays $1) than those for those with nearer dates, even if participants in both markets have the same average beliefs (Page and Clemen 2012). For example, consider a market trading contracts based on the outcome of the the roll of a 6-sided die. If the roll of the die was tomorrow, one would expect the price of "$1 if the die rolls 6" to be worth about $0.17, 1/6th of a dollar. But if the contracts were based on the exact same die being rolled 6 months from now, those contracts would sell for more than $0.17–perhaps $0.20. Conversely, contracts for "$1 if the die rolls 1–5" would not sell for $0.83, but for less–perhaps $0.80. Inferring a probability from the price (80% from $0.80) would therefore yield a biased prediction. As resolution extends into the future, predictions should become more and more biased towards 50%, and in consequence less and less informative.

This long-term bias toward 50% predictions might seem surprising, but it has been demonstrated both empirically and theoretically (D. Rothschild 2009; Antweiler 2012; Page and Clemen 2012). One reason for this is that for a long-term market, the potential purchaser of a prediction contract must decide between buying it and investing her money elsewhere, for example in the stock market. If she expects a 5% return on her money on the stock market, she will demand a 5% premium on her prediction contract. For a prediction she expects to come true 80% of the time, she should pay no more than $0.76 because if she paid more than $0.76 she would be better off investing that money in the stock market. If the same contract were for two years in the future, she should pay no more than $0.73. Similarly, successful prediction market traders might eschew long-term markets in favour of short-term markets where they can realize their trading advantage more quickly (Page and Clemen 2012, p. 512). Other reasons for investors to systematically avoid having their money tied up for long periods of time will have similar effects on prices. Further, as with the Foresight Exchange global warming market, long term prediction markets typically have very low trading volume (Page and Clemen 2012, p. 510), which makes it unlikely that their prices react correctly to new information. Therefore even if scientific predictions are amenable to controversy-free resolution, they may be unsuitable for prediction markets if their time horizon is too far in the future.[7] If long-term prediction markets are implemented, they might give a false impression of the probabilities underlying market prices.

Prediction market contracts must be based upon specific, resolvable predictions. While this poses no challenge for elections or sporting events where the object of interest *is* the result of a specific event, it poses a significant challenge

[7] Antweiler (2012) argues that the solution to this problem is "simple": compensate the holders of long-term prediction contracts with some combination of replicated investment portfolio and tax relief. His proposal might be simple in theory, but appears quite complex in practice. No currently operating prediction markets offer such compensation.

for science prediction market where predictions stand in for, or operationalize, competing scientific theories. While subjecting theoretical predictions to empirical test is central to the scientific ethos, interpreting the results of such tests is rarely straightforward. Popper pointed to Einstein's theory of general relativity as emblematic of good science, and Arthur Eddington's solar eclipse observations confirming Einstein's theory have often been invoked as a model of scientific testing. However, as Collins and Pinch (1993) relate, Eddington's measurements were anything but straightforward empirical tests of Einstein's theory. Many of Eddington's photographs were blurry; measuring the defection of starlight required comparing photographs taken many months apart; the measurements had to take into account the different characteristics of telescopes; the raw measurements gave conflicting results; and there was not universal agreement about how much deflection Einstein's theory predicted. Rather than an unambiguous measurement free of ideological bias, Eddington's results incorporated theories on multiple levels and required significant judgment. To what extent empirical testing of theories can be definitive is a matter of dispute. Kuhn (1996) argues that recalcitrant scientists can logically maintain their position in the face of an indefinite amount of apparently contrary evidence, while Galison (1997) argues that in practice competing scientists often agree about predictions and the implications of experiments. Nevertheless, interpreting the results of experiment is never unambiguous or value-free (Douglas 2009).

While scientific theories without any empirical consequences arouse justified suspicion, even theories with clear empirical consequences might not generate obvious predictions. Hanson (1995) claims that prediction markets might have hastened scientific agreement about continental drift. He envisions Wegener offering contracts "to be judged by some official body of geologists in a century" (Hanson 1995, p. 9), but this would not have been an effective criterion given the problem of 50%-bias in long-term prediction markets. Such a criterion is necessary, though, because it is unclear what resolvable predictions Wegener could have made at the time. His theory was based upon observations such as anatomical differences between Indonesian animals, the apparent drifting of Greenland by 11 meters per year, and the geological features of mountain ranges (Solomon 2001, pp. 89–90). All of these observations were already known, however, and thus couldn't be the basis of prediction contracts. His theory was not widely accepted until the 1960s, when a confluence of evidence including the discovery of volcanically active mid-ocean ridges convinced most scientists that the continents do indeed move (Solomon 2001, p. 104). If Wegener could have guessed that these ridges existed, he likely could have discovered them himself and profited from Hanson's proposal, but they were hardly an obvious consequence of his theory, and neither were any of the other observations that eventually led scientists to agree with his view. Prediction markets therefore would have been of little use to Wegener.

Prediction markets can transform the meaning of scientific theories in subtle ways that threaten to conflate theories with measurement operations. This applies even to theories that appear very amenable to prediction markets. The Foresight Exchange global warming prediction, for example, is judged by "glob-

ally averaged surface air temperature." While this may seem like a transparent measurement of global warming—it is after all measuring a globally averaged temperature—it is not. Firstly, the process of collecting global data in a form that allows for analysis and aggregation is not straightforward. Edwards (2010) documents the challenges of producing global data models of the climate. Not only are researchers geographically dispersed, but they are often separated by national and disciplinary boundaries. During the Cold War, for instance, it proved very difficult to get the meteorological services of different countries to agree even about how often throughout the day temperature measurements should be taken (Edwards 2010, p. 197). Training researchers in different countries to use the same methods can also be a challenge. Further, climate experimentalists, theoreticians, and weather forecasters all have different data requirements that can make sharing data difficult (Sundberg 2007). Even after all of these challenges have been overcome and data is collected and formatted in a standardized fashion, temperature measurement is not continuous. Rather, temperature measurements are made at particular locations and at particular times and then interpolated over the globe using computer models (Edwards 2011). Particularly in areas where measurement stations are sparse, such as in the middle of oceans, this can lead to a divergence between "measured" and true temperatures. The importance of this divergence is highlighted by the recent controversy over whether and to what extent there was a hiatus or slowing of global warming in the early 2000s (Tollefson 2014; Lewandowsky, Risbey, and Naomi Oreskes 2015). One explanation for the apparent slowing is "coverage bias": global average surface temperature is under-reported because areas such as the open oceans are incorrectly interpolated by current methods (Cowtan and Way 2013). Even if prediction markets correctly predict measured surface temperature, they might not predict actual surface temperature if the measured and actual surface temperatures diverge.

Secondly, surface air temperature is only a proxy measure for the temperature of the entire climactic system. The temperature of the upper atmosphere and the oceans may be equally as, or more important than, surface air temperature. Another possible explanation for the perceived slower than expected surface warming is that the "excess heat" was absorbed by the oceans (Balmaseda, Trenberth, and Källén 2013). If this explanation is correct, then globally averaged surface air temperature is a poor proxy for overall global temperature, and consequently prediction market prices based on surface air temperature could diverge from what they purport to predict: global warming. While this may be less of a problem for long time horizon markets as ocean heating begins to affect the overall climate, it could be quite significant for markets based upon near-term warming. Such markets would no longer effectively achieve a cognitive division of labor, as understanding the difference between surface air temperature and the temperature of the entire climate system would be required to correctly interpret the results of such markets.

Adopting large scale prediction markets could have significant unintended consequences for scientific practice. Since the early 1980s, historians and sociologists of science have observed significant changes in the organization of scien-

tific research, often characterized as the commercialization or commodification of science (Stephan 2012; Radder 2010). Sent and Mirowski (2008) describe this as a transition from a "Cold War" regime of science to a "globalized privatization regime." This era has been marked in the US by legal decisions, such as allowing the patenting of biological organisms and the patenting of publicly funded research, and by the flattening of federal funding for research alongside increased private financing. However, Kleinman (2010) argues that the changes to academic culture go far beyond what can be explained by legal and funding changes. While industry only funds about 7% of research according to Kleinmnan, this has led to "subtle and pervasive changes" in the culture of academic research (Kleinman 2010, pp. 24–25). Kleinman points to the anticipation of possible profit as a major reason for these changes: scientists focus on research with possible commercial applications in hope of future funding (Kleinman, Feinstein, and Downey 2012, p. 33). Along similar lines, Mirowski (2011) argues that patents and other commercialization-oriented legal products have proliferated in academia despite very few academic institutions turning a net profit from their intellectual property operations (Mirowski 2011, p. L2888).

Similarly outsized effects could be expected from a large-scale implementation of science prediction markets. If scientists anticipate that trading on prediction markets could generate significant profits, either due to being subsidized as Almenberg, Kittlitz, and Pfeiffer (2009) propose, or due to legal changes allowing significant amounts of money to be invested, they could shift their attention toward research that is amenable to prediction markets. The research most amenable to prediction markets is short-term and quantitative: the kind of research that is already encouraged by industry funding. Therefore prediction markets could reinforce an already troubling push toward short-term, application-oriented science. Further, scientists hoping to profit from these markets could withhold salient data in anticipation of using that data to make better informed trades than their peers. This could both slow publication and harm informal communication, and could even harm the accuracy of prediction markets themselves if accuracy depends upon data being made public, as Almenberg's experiments suggest.

Even if scientists do not pursue short term, easily resolvable research in pursuit of direct profits, prediction markets could distract from important scientific questions. If success in prediction markets is taken as a marker of scientific credibility, then scientists may pursue prediction-oriented research not to make direct profit, but to increase their reputation. The pursuit of "symbolic capital" (Bordieu 2004) might therefore have the same consequences as the pursuit of monetary capital.

The ability of prediction markets to produce transparent, reliable assessments of scientific claims is therefore largely illusory. The reliability of election prediction markets does not imply similar reliability for science prediction markets because science prediction markets are generally more longterm than election markets, because it is much more difficult to make resolvable predictions for science than for elections, and because any large-scale adoption of prediction markets would likely cause significant unintended harms to the organization of

scientific research.

5 Conclusion: Is the cure worse than the disease?

Current methods of assessing scientific knowledge are unsatisfactory. Peer review might not merely be a popularity contest, but there are reasons to doubt it as a reliable indicator of scientific merit. Various biases combined with a lack of inter-reviewer consistency mean that there is good reason to believe that the published record of science, circumscribed by the peer review process, does not accurately reflect the best available knowledge about questions of public interest. Consensus measures are an attractive solution to this problem because they provide a simple indicator of scientific opinion about questions that would otherwise be difficult for non-scientists to assess. However, consensus by itself does not indicate knowledge, and attempts by social epistemologists to distinguish epistemically justified from unjustified consensuses are not sufficient. Some fail to be applicable in practical circumstances, while others fail to be reliable or fail to maintain a cognitive division of labor, which eliminates the chief benefit of consensus measures. The problems with peer review and consensus can compound if consensus is judged based upon peer reviewed publications.

Prediction markets therefore have the potential to improve the operation of science. They could provide a transparent, reliable indicator of scientific knowledge about important questions while maintaining a cognitive division of labor. Although prediction markets are not infallible, they have proven reliable in limited domains, such as politics and sports. If science prediction markets were similarly reliable and posed no significant dangers to scientific practice, they would surely be worth aggressive implementation. Such markets, perhaps operated and subsidized by an entity such as the National Science Foundation, could be a boon to scientists and the public.

However, there are strong reasons to believe that science prediction markets would not perform as well as election or sporting prediction markets, and that they could pose significant dangers to scientific practice. Election and sporting markets have performed well at least in part because they are based on short-term, unambiguous predictions. Many scientific questions are not short-term and cannot be unambiguously resolved. Operationalizing scientific questions into predictions of quantitative measurements or definite observations risks answering questions that are subtly but importantly different from the questions they purport to answer. If interpreting the results of these markets requires detailed knowledge of the underlying subject, as is needed to distinguish global average surface air temperature from global average temperature, the division of cognitive labor promised by these markets will disappear. Perhaps worse, such predictions could be misinterpreted if people assume they accurately represent what they claim to. If prediction markets offered scientists expectations of profit, they could distort research priorities towards short-term, empirical ques-

tions, even if the overall profit potential of prediction markets is low. Even the *anticipation* of profit or the quest for reputation can have major consequences for the organization of scientific research.

Given these considerations, the promise prediction markets to solve problems in assessing scientific claims is largely illusory, while they could have significant unintended consequences for the organization of scientific research and the public perception of science. It would be unwise to pursue the adoption of prediction markets on a large scale, and even small scale markets such as the Foresight Exchange should be regarded with skepticism.

Nevertheless, prediction markets could be useful on a limited basis for some areas of science. Prediction markets perform best for easy to adjudicate predictions that can be resolved in the short term. Therefore their best uses could be for predicting the results of specific experiments, such as for the psychology reproducibility project (Dreber et al. 2015). Along similar lines, Potthoff (2007) has proposed using prediction markets to predict the results of clinical trials. Both of these applications are unaffected by the problems with prediction markets discussed in this paper. However, in neither of these cases are prediction markets expected to take the place of consensus as a measure of probable truth.

References

"A Social Epistemological Inquiry into Biases in Journal Peer Review" (n.d.). In: *Perspectives on Science*.

Almenberg, Johan, Ken Kittlitz, and Thomas Pfeiffer (2009). "An Experiment on Prediction Markets in Science". In: *PLoS ONE* 4.12, e8500.

Antweiler, Werner (2012). "Long-Term Prediction Markets". In: *Journal of Prediction Markets* 6.3.

Arrow, Kenneth et al. (2008). "ECONOMICS: The Promise of Prediction Markets". In: *Science* 320.5878, pp. 877–878.

Balmaseda, Magdalena A, Kevin E Trenberth, and Erland Källén (2013). "Distinctive climate signals in reanalysis of global ocean heat content". In: *Geophysical Research Letters* 40.9, pp. 1754–1759.

Bell, Tom W (2002). "Gambling for the good, trading for the future: The legality of markets in science claims". In: *Chapman Law Review*.

— (2006). "Prediction markets for promoting the progress of science and the useful arts". In: *George Mason Law Review* 14.37.

Biagioli, Mario (2002). "From Book Censorship to Academic Peer Review". In: *Emergences: Journal for the Study of Media & Composite Cultures* 12.1, pp. 11–45.

Biddle, Justin (2007). "Lessons from the Vioxx Debacle: What the Privatization of Science Can Teach Us About Social Epistemology". In: *Social Epistemology* 21.1, pp. 21–39.

Bordieu, Pierre (2004). *Science of Science and Reflexivity*. University of Chicago Press.

Carey, Benedict (2015). *Psychologists Welcome Analysis Casting Doubt on Their Work*. URL: http://www.nytimes.com/2015/08/29/science/psychologists-welcome-analysis-casting-doubt-on-their-work.html.

Collins, Harry and Trevor Pinch (1993). *The Golem: What Everyone Should Know About Science*. Cambridge University Press.

Cook, John, Dana Nuccitelli, et al. (2013). "Quantifying the consensus on anthropogenic global warming in the scientific literature". In: *Environmental ...*

Cook, John, Naomi Oreskes, et al. (2016). "Consensus on consensus: a synthesis of consensus estimates on human-caused global warming". In: *Environmental Research Letters* 11.4, pp. 1–7.

Cowtan, Kevin and Robert G Way (2013). "Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends". In: *Quarterly Journal of the Royal Meteorological Society*, n/a–n/a.

Douglas, Heather (2009). *Science, Policy, and the Value-Free Ideal*. Pittsburgh, Pa.: University of Pittsburgh Press.

— (2012). "Weighing Complex Evidence in a Democratic Society". In: *Kennedy Institute of Ethics Journal* 22.2, pp. 139–162.

Dreber, Anna et al. (2015). "Using prediction markets to estimate the reproducibility of scientific research". In: *Proceedings of the National Academy of Sciences of the United States of America* 112.50, pp. 15343–15347.

Edwards, Paul N (2010). *A Vast Machine: Computer Models, Climate Data, and The Politics of Global Warming*. MIT Press.

— (2011). "Representing the Global Atmosphere: Computer Models, Data, and Knowlede about Climate Change". In: pp. 1–18.

Fitzpatrick, Kathleen (2010). "Peer-to-peer Review and the Future of Scholarly Authority". In: *Social Epistemology* 24.3, pp. 161–179.

Forsythe, Robert et al. (1992). "Anatomy of an experimental political stock market". In: *The American Economic Review*, pp. 1142–1161.

Galison, Peter (1997). *Image and Logic: A Material Culture of Microphysics*. University of Chicago Press.

Gilbert, D T et al. (2016). "Comment on "Estimating the reproducibility of psychological science"". In: *Science* 351.6277, pp. 1037–1037.

Hanson, Robin (1995). "Could gambling save science? Encouraging an honest consensus". In: *Social Epistemology* 9.1, pp. 3–33.

HERI (2014). "Undergraduate Teaching Faculty: The 2013-2014 HERI Faculty Survey". In: pp. 1–117.

Horrobin, David F (1990). "The philosophical basis of peer review and the suppression of innovation". In: *JAMA: the journal of the American Medical Association* 263.10, pp. 1438–1441.

Hsu, Shi-Ling (2011). "A Prediction Market for Climate Outcomes". In: *University of Colorado Law Review*.

Jacobs, Peter (2014). "It Ain't Just the Heat, It's the Humanity:fffdfffdfffdEvidence and Implications of a Knowledge-Based Consensus on Climate Change". In: *AGU Science Policy Conference*.

Kleinman, Daniel Lee (2010). "The Commercialization of Academic Culture and the Future of the University". In: *The Commodification of Academic Research: Science and the Modern University*. Ed. by Hans Radder. University of Pittsburgh Press, pp. 1–21.

Kleinman, Daniel Lee, Noah Weeth Feinstein, and Greg Downey (2012). "Beyond Commercialization: Science, Higher Education and the Culture of Neoliberalism". In: *Science & Education* 22.10, pp. 2385–2401.

Knorr-Cetina, K (1999). *Epistemic cultures : how the sciences make knowledge*. Cambridge, Mass.: Harvard University Press.

Kuhn, Thomas S (1996). *The Structure of Scientific Revolutions*. University of Chicago Press.

Lee, Carole J et al. (2012). "Bias in peer review". In: *Journal of the American Society for Information Science and Technology* 64.1, pp. 2–17.

Lewandowsky, Stephan, James S Risbey, and Naomi Oreskes (2015). "On the definition and identifiability of the alleged "hiatus" in global warming". In: *Nature Publishing Group* 5, pp. 1–13.

Longino, Helen E (1994). "Knowledge in Social Theories of Science". In: *Socializing Epistemology*, pp. 135–157.

Miller, Boaz (2012). "When is consensus knowledge based? Distinguishing shared knowledge from mere agreement". In: *Synthese* 190.7, pp. 1293–1316.

Mirowski, Philip (2011). *Science-Mart: Privatizing American Science*. Kindle Edition. Harvard University Press.

Nicholson, Joshua M and John PA Ioannidis (2012). "Conform and be funded". In: *Nature* 492, pp. 34–36.

Open Science Collaboration (2015). "PSYCHOLOGY. Estimating the reproducibility of psychological science." In: *Science* 349.6251, aac4716–aac4716.

Oreskes, N (2004). "BEYOND THE IVORY TOWER: The Scientific Consensus on Climate Change". In: *Science* 306.5702, pp. 1686–1686.

Page, Lionel and Robert T Clemen (2012). "Do Prediction Markets Produce Well-Calibrated Probability Forecasts? *". In: *The Economic Journal* 123.568, pp. 491–513.

Pennock, David M et al. (2001). "The real power of artificial markets". In: *Science* 291.5506, pp. 987–988.

Peters, Douglas P and Stephen J Ceci (1982). "Peer-review practices of psychological journals: The fate of published articles, submitted again". In: *Behavioral and Brain Sciences* 5.02, pp. 187–195.

Pfeiffer, Thomas and Johan Almenberg (2010). "Prediction markets and their potential role in biomedical research–A review". In: *Biosystems* 102.2, pp. 71–76.

Pinto, Manuela Fernández (2014). "Philosophy of science for globalized privatization". In: *Studies in History and Philosophy of Science* 47.C, pp. 10–17.

Potthoff, Richard F (2007). "Prediction markets, Bayesian priors, and clinical trials". In: *Journal of Statistical Planning and Inference* 137.11, pp. 3706–3721.

Radder, Hans (2010). "The Commodification of Academic Research". In: *The Commodification of Academic Research: Science and the Modern University*. Ed. by Hans Radder. University of Pittsburgh Press, pp. 1–24.

Rosenberg, Stacy et al. (2009). "Climate change: a profile of US climate scientists' perspectives". In: *Climatic Change* 101.3-4, pp. 311–329.

Rothschild, D M and R Sethi (2015). "Wishful Thinking, Manipulation, and the Wisdom of Crowds: Evidence from a Political Betting Market".

Rothschild, David (2009). "Forecasting Elections: Comparing Prediction Markets, Polls, and Their Biases". In: *Public Opinion Quarterly* 73.5, pp. 895–916.

Rothwell, Peter M and Christopher N Martyn (2000). "Reproducibility of peer review in clinical neuroscience Is agreement between reviewers any greater than would be expected by chance alone?" In: *Brain* 123.9, pp. 1964–1969.

Sent, Esther-Mirjam and Philip Mirowski (2008). "The Commercialization of Science and the Response of STS". In: *The Handbook of Science and Technology Studies*. Ed. by Edward J. Hackett and Society for Social Studies of Science. Cambridge, Mass.: MIT Press.

Smith, Richard (2006). "Peer review: a flawed process at the heart of science and journals". In: *Journal of the royal society of medicine* 99.4, pp. 178–182.

Solomon, Miriam (2001). *Social Empiricism*. Cambridge, Mass.: MIT Press.

Stegenga, Jacob (2011). "An impossibility theorem for amalgamating evidence". In: *Synthese* 190.12, pp. 2391–2411.

Stephan, Paula (2012). *How Economics Shapes Science*. Harvard University Press.

Sundberg, M (2007). "Parameterizations as Boundary Objects on the Climate Arena". In: *Social Studies of Science* 37.3, pp. 473–488.

Sunstein, Cass R (2006). *Infotopia: How Many Minds Produce Knowledge*. Oxford University Press.

Tollefson, Jeff (2014). "Climate change: The case of the missing heat". In: *Nature* 505.7483, pp. 261–262.

Tucker, Aviezer (2003). "The epistemic significance of consensus". In: *Inquiry* 46.4, pp. 501–521.

Who said Brexit was a surprise? (2016). URL: `http://www.economist.com/blogs/graphicdetail/2016/06/polls-versus-prediction-markets`.

Wolfers, Justin and Eric Zitzewitz (2004). "Prediction Markets". In: *Journal of Economic Perspectives* 18.2, pp. 107–126.

Ziman, John (2004). *Real Science: What it is, and what it means*. What it is, and what it means. Cambridge University Press.

Zuckerman, Harriet and Robert K Merton (1971). "Patterns of evaluation in science: Institutionalisation, structure and functions of the referee system". In: *Minerva* 9.1, pp. 66–100.