University of California, Irvine
Department of Logic and Philosophy of Science

Essays in the Epistemology of Science

Doctoral Dissertation of:
Aydin Mohseni

Advisors:
Prof. Cailin O'Connor
Prof. Simon Huttegger

June 2021

Acknowledgments

I must thank my gracious advisors, Cailin O'Connor and Simon Huttegger, for their priceless counsel and company. I have learned enormously from them both. I've also received guidance and support from several of the faculty here at LPS at UCI and would like to express heartfelt thanks in particular to Brian Skyrms, Kyle Stanford, Jeff Barrett, and Jim Weatherall. Thank you. I would be remiss if I failed to thank my brilliant peers, Daniel Herrmann, Gerard Rothfus, Cole Williams, Nikhil Addleman, and Gabe Orona for their companionship in philosophical co-adventure. Finally, my thanks go to my wonderful parents who have made this life both possible and beautiful.

Aydin Mohseni
Irvine
June 2021

Abstract

What is the right statistical theory for scientific practice? What are the distinctive dynamics of social learning and conformity? And what can disagreement between theoretical models teach us? This dissertation consists of essays that engage just these questions.

Chapter 1 examines explanations for a broadly maligned research practice, HARKing, which has been called one of the 'four horsemen' of the replication crisis in the social and biomedical sciences (alongside publication bias, low statistical power, and p-hacking). In it, I demonstrate how classical accounts of why HARKing undermines the reliability of scientific findings must be wrong, and proffer what I take to be the correct, Bayesian account for when and why HARKing is in fact bad. Further, I consider the implications of all this for a prominent proposal for methodological reform in the context of the replication crisis.

Chapter 2 consists of more exploratory work, produced in collaboration with my colleague Cole Williams. We propose a simple model of social learning on networks under the influence of conformity bias. In our model, heterogeneous agents express public opinions where those expressions are driven by the competing priorities of accuracy and of conformity to one's peers. Agents learn, by Bayesian conditionalization, from private evidence from nature, and from the public declarations of other agents. Our key findings are that networks that produce configurations of social relationships that sustain a diversity of opinions empower honest communication and more accurate beliefs and that the networks that do this best turn out to be those which are both less centralized and less connected.

Finally, chapter 3 consists of elucidating the relationship between two key models in evolutionary game theory–the replicator dynamics and Moran process. These models are connected by a mean-field relationship—the former describes the expected behavior of the latter. However, there are conditions under which their predictions diverge. I demonstrate that the divergence between their predictions is a function of standard techniques used in their analysis, and of differences in the idealizations involved in each. My analysis reveals problems for stochastic stability analysis in a broad class of games. I demonstrate a novel domain of agreement between the dynamics and consider a simple moral for scientific modeling.

Contents

HARKing: from Misdiagnosis to Misprescription 1

Abstract

The practice of HARKing—hypothesizing after results are known—is commonly maligned as undermining the reliability of scientific findings. There are several accounts in the literature as to why HARKing undermines the reliability of findings. We argue that none of these is right and that the correct account is a Bayesian one. HARKing can indeed decrease the reliability of scientific findings, but it can also increase it. Which effect HARKing produces depends on the difference of the prior odds of hypotheses characteristically selected ex ante and ex post to observing data. Further, we show how misdiagnosis of HARKing can lead to misprescription in the context of the replication crisis.

1.1 Introduction

In a 2019 article in *Nature*, the author, psychologist Dorothy Bishop, describes HARKing as one of "the four horsemen of the reproducibility apocalypse," along with publication bias, low statistical power, and *p*-hacking (Bishop, 2019, p. 435). The practice of HARKing—hypothesizing after results are known—is commonly maligned as undermining the reliability of scientific findings.[1] There are several accounts in the literature as to why HARKing undermines the reliability of findings. Scholars have argued that HARKing undermines frequentist guarantees

[1] See, for example, Kerr (1998); John et al. (2012); Rubin (2017); and Murphy and Aguinis (2019).

of long-run error control, that it violates a broadly Popperian picture of science, and misrepresents hypotheses formulated ex post to observing the data as those formulated ex ante. We argue that none of these accounts correctly identify *why* HARKing can undermine the reliability of findings, and that the correct account is a Bayesian one.

We will show that HARKing can indeed decrease the reliability of scientific findings, but that there are conditions under which HARKing can actually increase the reliability of findings. In both cases, the effect of HARKing on the reliability of findings is determined by the difference of the prior odds of hypotheses characteristically selected ex ante and ex post to observing data. To make this precise, we employ a standard model of null hypothesis significance testing in which we provide necessary and sufficient conditions for HARKing to decrease the reliability of scientific findings.

The aim of this paper is not to defend the practice of HARKing. Insofar as HARKing involves disclosing less than complete information to those who wish to learn and act based on scientific findings, it is clearly epistemically undesirable.[2] HARKing can also be ethically and pedagogically undesirable, insofar as it involves intentional deception or presenting an inaccurate model of science to students. Rather, the aim here is to clarify the relationship between HARKing and the reliability of scientific findings.

Understanding HARKing is important on at least two counts. Historically, HARKing is closely tied to questions regarding the relationship between prediction and accommodation. These questions have engaged philosophers at least as early as Mill (1843), were made central in the philosophy of science by Popper (1934) and continue to be of concern in contemporary discussions in scientific epistemology.[3] As mentioned, HARKing is also imputed to be among the questionable research practices contributing to the crisis of replication in the social and biomedical sciences, which has rightly become a subject of interest to philosophers of science.[4] A better understanding of HARKing sheds light on both these issues.

The strategy for demonstrating that standard accounts for why HARKing leads to unreliable findings are incorrect is as follows. Each account, i, claims that

[2] For precise, decision-theoretic formulations of this observation see the value of knowledge theorems of Savage (1954), Good (1967), and Skyrms (1990, Ch. 4).

[3] See Hitchcock and Sober (2004); Douglas and Magnus (2013); Mayo (2014); Barnes (2014); Worrall (2014) and (Schurz, 2014).

[4] For excellent philosophical examinations of social and epistemic issues involved in the replication crisis see Romero (2019, 2020); Romero and Sprenger (2020); Heesen (2018); Bruner and Holman (2019); Bright (2017); Bird (2020); Devezer et al. (2019); Baumgaertner et al. (2019); and Machery (2020).

HARKing undermines the reliability of scientific findings because HARKing exhibits a particular property, φ_i, distinct to that account. We show that HARKing can increase the reliability of findings while still satisfying property φ_i and, hence, φ_i cannot explain why HARKing is in fact bad for the reliability of findings. Instead, we provide a Bayesian analysis of HARKing that provides necessary and sufficient conditions for when HARKing worsens or improves the reliability of findings.

In §1.2, we summarize several accounts of why HARKing is bad for the reliability of scientific findings. In §1.3, we present clear criteria for the reliability of scientific findings with which to measure the effect of HARKing relative to specific alternatives. In §1.4, we present a standard model of hypothesis testing with which to reason about the statistical consequences of HARKing. In §1.5, we provide necessary and sufficient conditions for when HARKing improves and worsens the reliability of findings. In §1.6, we show how misdiagnosis of HARKing ramifies into misguided proposals for redefining statistical significance in the context of the replication crisis. In §1.7, we conclude with a discussion.

1.2 HARK! Who Goes There?

First, let us be clear about what we mean here by HARKing. The term 'HARKing' was first coined by social psychologist, Norbert Kerr, in his 1998 article "HARKing: hypothesizing after results are known." Kerr defines HARKing as "...presenting a post hoc hypothesis in the ... [study] report as if it were an a priori hypothesis" (Kerr, 1998, p. 197).[5] HARKing occurs when a researcher selects her study hypothesis after observing the data and reports this hypothesis as if it had been formulated prior to observing the data—that is, as if it had been a prediction. This is typically contrasted with the normative protocol in which the researcher selects a hypothesis prior to observing the data, and then, after observing her data, reports whether the hypothesis attained significance given some conventional threshold for statistical significance.

In his 1998 article, Kerr anticipates many of the now-standard objections to the practice of HARKing. These include taking unjustified statistical license, propounding theories that cannot pass Popper's falsifiability criterion, and disguising post hoc explanations as a priori explanations (Kerr, 1998, p. 211). Since then, and especially in light of the replication crisis, philosophers of

[5] Kerr employs 'a priori' and 'a posteriori' to mean before and after the event of observing one's study data. We use the terms 'ex ante' and 'ex post' for these to avoid confusion with the standard philosophical meanings of the former terms.

science and scholars in the social and biomedical sciences have elaborated and propounded these accounts.[6]

Several variants of HARKing exist in the literature and should be distinguished. STARKing, or story telling after results are known, is where a finding is presented along with a narrative produced ex post to observing the data meant to bolster the plausibility of that finding. THARKing, or transparent hypothesizing after results are known, is where it is clearly communicated that the study hypothesis was selected after the data were observed (Hollenbeck and Wright, 2016). We concern ourselves here only with HARKing. In particular, we are concerned with accounts of the epistemic effect of HARKing: why, precisely, it undermines the *reliability* of scientific findings, as presented in (Kerr, 1998) and in other influential accounts such as (Rubin, 2017) and (Mayo, 2019).

HARKing as Undermining Error Control

The first account of the epistemic problem of HARKing emerges straightforwardly from a classical, frequentist philosophy of statistics that concerns itself with error-control. In the context of hypothesis testing, a central strand of frequentist thought locates the reliability of tests in terms of their guarantees of controlling the long run frequencies of Type I and Type II error in hypothetical repetitions of those tests (Lehmann, 1993).

An example of such a guarantee is as follows. Consider a hypothesis test with a conventional significance threshold of $\alpha \in [0, 1]$ (corresponding to its Type I error rate). A sample of data is collected for which a test statistic, t, is determined.[7] A decision to reject or fail to reject the null hypothesis is made as follows. On the assumption that the null hypothesis is true,[8] one determines the p-value for the test, or the probability of having observed a test statistic at least as extreme was actually observed, $p = P(T > t|H_0)$. If this value meets the significance threshold, $p < \alpha$, then the null hypothesis is rejected. If the threshold is not met, one fails to reject the null. In a world where the null hypothesis is true, such a test produces mistaken rejections of the null hypothesis $100 \times \alpha$ percent of the time if the test were repeated infinitely many times.[9]

[6] See, in particular, Leung (2011); Rubin (2017, 2019) and Hollenbeck and Wright (2016).

[7] For example, the test statistic may be the mean of the difference in, or association between, two variables in a data set.

[8] And also that the inductive assumptions of the test hold true—e.g., normality, homoscedasticity, probabilistic independence, and so on.

[9] More precisely, this occurs *almost surely* with respect to the measure over the infinite sequence of outcomes.

HARKing undermines such guarantees. When a researcher engages in HARKing, she waits until after she observes her data and then selects a hypothesis to report from among those that are statistically significant. To drive our point home, consider a researcher who has infinitely many probabilistically independent hypotheses from which she may choose. Further, imagine that all of her hypotheses are false. For any positive Type I error rate, $\alpha > 0$, she will obtain statistically significant results, as mistaken rejections of the null are now a certainty. If she engages in HARKing then she will only ever report significant findings, even though all of her candidate hypotheses are false, and so guarantees of error-control of the sort just described become ill-defined.

Consider the following formulation of the problem by Rubin: "For example, if a researcher tests 20 hypotheses with an alpha level of .05, then he has a 64.15% chance of making at least one Type I error. However, if his results confirm only one of these hypotheses, and he decides to suppress the other 19 disconfirmed hypotheses, then he will give the incorrect impression that he only conducted a single hypothesis test and that, consequently, he only had a 5% chance of making a Type I error" (2017, p. 14). Bishop echoes a familiar refrain in describing the consequences of HARKing: "P-values are meaningless when taken out of context of all the analyses performed to get them" (Bishop, 2019, p. 435).

This is indeed correct: HARKing undermines frequentist guarantees of long run error control. However, we are interested in the *reliability* of scientific findings, and the types of error that the frequentist promises to control—i.e., Type I and Type II error rates—simply do not capture the reliability of findings. The Type I and II error rates of a test tell us that *if* the hypothesis is true or false, *then* what long run frequency of errors is to be expected.[10] We discuss this further in §1.3 where we provide a natural and practicable measure of the reliability of findings.

HARKing as Failing to Provide a Severe Test

The second account maligns HARKing for violating a broadly Popperian picture of science. The basic idea is that when a hypothesis is selected ex post to observing data for its compatibility with those data, then it could not be reliably disconfirmed by those data. This is described by Rubin as the objection that HARKing is "problematic for scientific progress because it results in hypotheses that are always confirmed and never falsified by the results." (Rubin, 2017,

[10] And the p-value tells us the probability, conditional on the null being true, of observing data at least as extreme as was actually observed.

p. 2) Kerr states this plainly: “HARKed hypotheses fail Popper’s criterion of disconfirmability” (1998, 205).

That said, we need to sharpen this objection, as *deductive* falsification is typically not feasible for statistical tests of hypotheses. The most sophisticated version of this view which has risen to prominence in the philosophy of statistics is the severe testing account propounded by Deborah Mayo (Mayo, 2019, 2006). On this account, a statistical test provides us with corroboration of a hypothesis insofar as it submits that hypothesis to a severe test, where a severe test is one that would reliably detect an error in that hypothesis if one were present (Mayo and Spanos, 2006, 2011). This is Popper’s criterion of falsifiability adapted to the statistical context.

It is clear why a hypothesis reported via HARKing fails to satisfy the requirements of a severe test. Recall that under HARKing the researcher selects her hypothesis after observing her data from among the set of hypotheses that are significant given those data. The reported results of her study are significant by construction. She could have failed to report the hypothesis that she in fact did, but she would have not have reported its ‘falsification’ by the data if it had turned out to be non-significant. Thus, HARKing fails to provide warrant for belief in hypotheses because it fails to expose them to opportunities for statistical falsification.

As with frequentist error rates, the relationship between the severity of tests and the reliability of scientific findings is not a direct one. What we will see is that HARKing can fail both to provide well-defined frequentist error rates or meet the requirements of severe testing, and yet produce more reliable findings, and so, whatever virtues these accounts capture, they fail to explain the interaction of HARKing with the reliability of scientific findings.

HARKing as Misrepresentation

Our third account, from (Kerr, 1998), posits that HARKing misrepresents hypotheses formulated ex post to observing the data as those formulated ex ante. As mentioned, we have no truck with the version of this objection that locates the ill of HARKing in ethical terms. But this objection rarely goes beyond observing the fact of misrepresentation to identifying why, precisely, misrepresentation of this sort should negatively effect the reliability of findings. That it does so appears to be just assumed.

Our analysis will explain the relationship between HARKing and the reliability of findings. In particular, we will show that the last objection is closest to the mark: it is misrepresentation of the hypotheses that can be epistemically detrimental. When HARKing is bad, it is because it can lead us to mistakenly infer that a hypothesis reported via HARKing enjoyed greater ex ante evidential support than it in fact did, and so garners greater ex post confidence than it in fact deserves. Let us to turn to our Bayesian account.

1.3 The Reliability of Scientific Findings

Our argument requires an adequate specification of the reliability of scientific findings with which to determine whether a given account of HARKing correctly diagnoses the effect of HARKing on the reliability of findings. For this, we draw on existing statistical concepts—specifically false discovery and false omission rates—to formalize a natural notion of reliability in the context of null hypothesis significance testing. We argue for why this is a more apt characterization of reliability than frequentist guarantees of error-control, specifically, Type I and II error rates, or data-dependent notions such as the 'severity' of a test.

Classically, we want our epistemic methods to produce fewer false beliefs and more true ones. In the statistical context, we can ask that fewer of the results we declare significant be false and fewer of the results we declare non-significant be true. These correspond to the requirement that our methods exhibit lower rates of false discovery and false omission respectively.[11]

The *false discovery rate* (FDR) of a population of studies is the expected proportion of its findings (rejections of the null hypothesis) that are false findings (where the null hypothesis is true). For a given method of reporting findings, M, and population of studies, the FDR corresponds to:

$$FDR(M) = Pr(H_0 \mid \text{significant}; M) = \frac{Pr(H_0, \text{significant}; M)}{Pr(\text{significant}; M)}.^{12}$$

That is, the ratio of false and significant findings over over all significant findings. Intuitively, the reliability of research increases as the false discovery rate decreases. Indeed, in the context of the replication crisis, the rate at which findings in a literature fail to replicate under more stringent tests is an estimator for the false discovery rate of that literature.

[11] The notion of the false discovery rate of studies was first introduced to the statistical sciences by Bejamin and Hochberg (1995) for explicating the expected frequencies of true and false hypotheses in the context of multiple testing.

[12] Note that the probability of individual study outcomes is employed here as it is mathematically equivalent to the corresponding expected fraction of study outcomes. For example, the probability of a single study hypothesis obtaining significance while the null is true is equivalent to the expected fraction of many study hypotheses that obtain significance while their nulls are true.

The *false omission rate* (FOR) of a literature is the expected proportion of its negative findings (failures to reject the null hypothesis) that are false negative findings (where the alternative hypothesis is true). For a given method of reporting findings, M, and population of studies, the false omission rate corresponds to:

$$FOR(M) = Pr(H_1 \mid \text{not significant}; M) = \frac{Pr(H_1, \text{not significant}; M)}{Pr(\text{not significant}; M)}.$$

That is, the ratio of true and non-significant findings over all non-significant findings. Intuitively, the reliability of studies increases whenever their false omission rate decreases. In particular, the false omission rate provides a measure of how poorly research detects the truth of hypotheses.

Protocols for selecting hypotheses and reporting results can affect both the false discovery and omission rates of a population of studies. For our analysis, we consider broadly two sorts of protocols: one which requires the researcher to commit to a hypothesis prior to observing the data, and another that allows her to select her hypothesis only after observing the data. Some changes in methods increase one type of error while decreasing the other.[13]

Definition 1 (Reliability of a Method)**.** Let a population of studies be specified by the significance threshold, $\alpha \in (0, 1)$, and mean power,[14] $1 - \beta \in (0, 1)$, of its tests along with the prevalence of true hypothesis, $\pi \in (0, 1)$, among the set of hypotheses selected for testing.

Say that a given reporting method, M, is *more reliable* with respect to *false discovery* than another, M', if, for any such population of studies, findings produced under M yield a lower false discovery rate than those under M'.

$$FDR(M) \leq FDR(M'), \quad \text{for any } \pi, \alpha, \beta \in (0, 1).$$

Similarly, say that M is *more reliable* with respect to *false omission* than M' if, for any such population of studies, findings produced under M produce a lower false omission rate than those under M'.

$$FOR(M) \leq FOR(M'), \quad \text{for any } \pi, \alpha, \beta \in (0, 1).$$

[13] For example, it is well-known that lowering the threshold for statistical significance will tend to decrease false discovery while increasing false omission.

[14] Statistical power is the complement of Type II error. That is, the power of a test is the probability, conditional on the alternative hypothesis being true, of correctly rejecting the null hypothesis.

Note that we require that there is some non-zero fraction of true and false hypotheses to be tested—that is, $\pi \in (0, 1)$. If this were not so, improvements in method could not improve the FDR and FOR of studies.[15]

These provide natural, practical measures of reliability. But what of classical Type I and II error rates? To see why Type I and II error rates cannot capture the reliability of findings, consider a domain of scientific study in which no true hypotheses are available—that is, where the null hypothesis is always true. Regardless of the Type I and II error rates of tests, all significant findings will be false findings. More generally, one and the same frequentist guarantee—Type I and II error rates—is compatible with any reliability of scientific findings since reliability is critically a function of the prevalence of true hypotheses put to test.

Other measures of reliability might be sought in data-dependent measures, such as p-values or the severity of a test.[16] Yet the same shortcoming applies to such measures. Insofar as a measure ignores the prevalence of true hypotheses submitted to testing, it will assign the same measure of reliability to a set of findings assured to be false as to a set of findings assured to be true.

In the context of classical null hypothesis significance testing, what we may want from a measure of reliability of scientific findings is that more of those claimed to be statistically significant are in fact true and more of those claimed not to be significant are in fact false.[17] The false discovery and omission rates of hypotheses capture just this; and with them in hand, we can proceed to an analysis of the interaction of HARKing and the reliability of scientific findings.

It might be argued that the prevalence of true hypotheses is not a quantity generally known to us, especially given the distortions produced by publication bias, so a measure of reliability that uses this unknown quantity is useless. There are two problems with this objection. First, reasonable estimates of the prevalence of true hypotheses in a domain are difficult but not impossible to produce (Dreber et al., 2015). Second, and more to the point, we can still entertain the hypothetical: we can ask what the false discovery and omission rates

[15] Though changes in method could improve one of them. For example, if all candidate hypotheses are true, a method could not improve the FDR (since there could be no false discoveries), but it could improve the FOR merely by assigning significance to more results.

[16] The severity of a test can be thought of as data-dependent analogue of statistical power (Mayo-Wilson and Fletcher, 2019).

[17] Alternatively, one may wish to move to leave the NHST paradigm for, for example, a fully Bayesian approach to analyzing scientific findings in which one applies credences over the truth of one's study hypotheses. The authors cautiously endorse proposals of this sort (see, for example, Etz and Vandekerckhove (2016) and Romero and Sprenger (2020)), but recognize that it is worthwhile improving existing statistical practices even as we work toward more substantive, long-term changes in method.

of different reporting protocols *would be* given different underlying prevalences of true hypotheses submitted for testing. And we can learn if one method outperforms another *regardless* of whether truth is a rare disease or as common as pig tracks.

1.4 When HARKing Cannot be Bad

Consider a world in which all study hypotheses have equal prior odds. We will show that in such a world HARKing cannot lessen the reliability of findings. Let us see why this is so and consider the implications of this fact.

First, a note on how to interpret the prior odds of hypotheses. In our model,[18] the prior odds of a hypothesis are to be understood in terms of a well-defined prevalence of true study hypotheses. A study hypothesis H_i belongs to a set of candidate hypotheses, $\boldsymbol{H} = \{H_i\}_{j=1}^{n}$, from which it is selected by the researcher. A given fraction of the hypotheses in the set, $\pi \in (0, 1)$, are true and their complements false. If hypotheses are randomly selected from this set for testing, the prior probability of a study hypothesis is just the probability of selecting a true hypothesis from this set $Pr(H_i) = \pi$.[19,20]

Now, let us define our research methods. In the endorsed picture of hypothesis testing, the researcher selects her hypothesis, H_i, from the set of possible hypotheses prior to observing her data. Only then does she observe her data, and then she reports whether her predicted hypothesis, H_i, was statistically significant given the conventional threshold for significance, α. Call this protocol *prediction* and denote it M^p.

In contrast, under a protocol of *HARKing* the researcher first observes her data, and then selects a hypothesis H_i at random from the set of hypotheses that have turned out to be statistically significant in light of her data $\{H_i \in \boldsymbol{H} | p_i < \alpha\}$, if the set is nonempty. Denote this protocol M^h.

We summarize the preceding two reporting protocols as follows.

[18] This is an adaption of a famous model of the reliability of findings in science (Ioannidis, 2005; Maniadis et al., 2014).

[19] Equivalently, the prior odds of the hypothesis will be $1 : (\pi^{-1} - 1)$. We use the terms 'prior odds' and 'prior probabilities' to denote the same quantity.

[20] For the Bayesian, the analysis is more straightforward: the prevalence of true hypotheses is just her prior. The stipulation of a process of random selection of hypotheses is provided to make the analysis more palatable to an interlocutor skeptical of ostensible subjectivity of the Bayesian approach; priors here correspond to objective elements of the probability model—fractions of true hypotheses in a well-defined population of hypotheses—and not subjective beliefs.

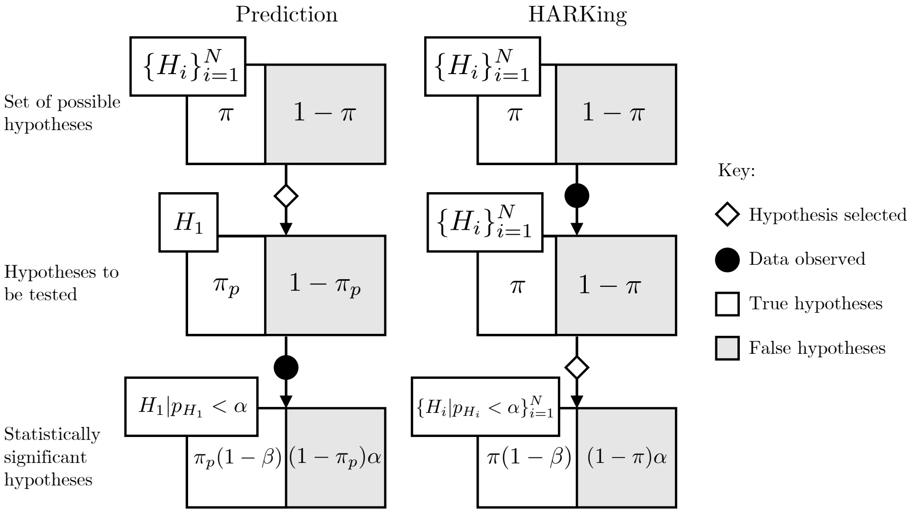


Figure 1.1: The filtration of study hypotheses for reporting via prediction and HARKing protocols.

1. *Prediction M^p*: Prior to observing the data, select a single hypothesis to test. Report the hypothesis if turns out to be significant.
2. *HARKing M^h*: After observing the data, randomly select a hypothesis from among those that are significant (if there are any) and report the hypothesis.

Note that we assume that a hypothesis is reported only if it is significant. This reflects the reality of publication bias and the concomitant file drawer effect. That said, nothing critical turns on this assumption, and later we will allow for some fraction of statistically non-significant findings to be reported as well when we consider the false omission rates of protocols.

Given our specification of these protocols, we can demonstrate the following. (All proofs are provided in the mathematical appendix.)

Proposition 1. When hypotheses are selected from the same set of candidate hypotheses with fraction $\pi \in (0,1)$ true hypotheses, then prediction yields the same false discovery rate as HARKing.

$$FDR(M^p) = FDR(M^h)$$

That is, in such a case, HARKing is as reliable with respect to false discovery as prediction. The logic of the result is simple. By stipulating that hypotheses be selected at random, we have set the fraction of true hypotheses selected via both prediction and via HARKing to be equal. And, in both cases, selected hypotheses are reported only if they are significant. Thus, the fraction of hypotheses that are true as well as significant is identical. See figure 1.1, where the only effectual difference between the protocols is that prediction checks only a single hypothesis at a time, whereas HARKing checks all candidate hypotheses; both methods produce identical statistics in reported hypotheses. The difference is that a researcher employing HARKing filters a larger set of hypotheses for statistical significance. In this world, the researcher employing HARKing is simply more efficient in filtering hypotheses for significance—while the fraction of her true and false discoveries is the same, her absolute rate of discovery is strictly greater.

What of false omission rates? Missing out on true hypotheses is not as often voiced as a concern. The worries around HARKing typically focus on its contribution to high false discovery rates, which correspond to low replication rates. However, under the same assumptions—hypotheses submitted to both protocols exhibit equal prior odds—HARKing performs no worse than prediction with respect to false omission either. This fact follows the same reasoning as the equality of false discovery rates (and is proved in the mathematical appendix). Both prediction and HARKing protocols filter hypotheses with the same frequencies, only HARKing does so more efficiently.

Let us revisit an assumption. When engaging in HARKing and confronted with multiple significant hypotheses, we assumed that the researcher selects one at random. What if, instead, she reports the most impressive, publication-worthy result—the result of the lowest p-value? This natural description yields the following reporting protocol.

3. *Selective HARKing* M^{sh}: After observing the data, select the hypothesis with the lowest p-value from among those that are significant (if there are any) and report the hypothesis.

Hypotheses that yield lower p-values are, on average, more likely to be true. Thus, the population of studies reported via selective HARKing will be composed of more true findings than either of those produced via prediction or HARKing. And as more hypotheses are considered, the lower the expected p-value of the hypothesis with the least p-value will be, and thus the fraction of findings consisting of true discoveries will be greater.

Proposition 2. When hypotheses are selected from the same set of possible hypotheses with fraction $\pi \in (0, 1)$ true hypotheses, and there are more than two candidate hypotheses, then:

1. Prediction yields a false discovery rate exceeding that of selective HARKing, $FDR(M^p) > FDR(M^{sh})$.
2. The false discovery rate for selective HARKing is decreasing in the size of the set of candidate hypotheses L.

That is, a slightly more sophisticated version of HARKing can produce strictly and substantially more reliable findings than prediction.

1.5 When HARKing Must be Bad

When is HARKing bad for the reliability of scientific findings? Under the conditions we described—equal prior odds of hypotheses selected via prediction and HARKing—selecting hypotheses ex post to observing the data cannot undermine the reliability of one's findings. Thus, for HARKing to be bad in the course of normal scientific practice, one of the assumptions of our model must not obtain. The natural candidate is that scientists do not in fact choose their hypotheses at random.

A researcher's choice of study hypothesis tends to be informed by her domain knowledge. The hypotheses she chooses prior to observing the data may be informed by theory, previous findings in the literature, and common sense. We can imagine that, when a researcher is suitably informed, a hypothesis she selects is more likely to be true than a hypothesis selected at random from the set of hypotheses that are merely logically consistent with her data. In such a case, we can expect the researcher to do better than chance prior to observing the data. Under prediction, such hypotheses are then filtered by their statistical significance after the data have been observed.

When a researcher is engaged in HARKing, however, she is no longer filtering the set of possible hypotheses via her domain knowledge. Rather, the full set of hypotheses consistent with her data are submitted to the filter of statistical significance.

Formally, this corresponds to a model in which the researcher chooses from two sets of hypotheses with different fractions of true and false hypotheses. When she chooses ex ante to observing her data, via prediction, she tests a subset of hypothesis with some prior odds; and when she chooses ex post to observing her data, via HARKing, she chooses from a superset with different prior odds. For HARKing to be bad in such a world, it *must* be the case that the prevalence of true hypotheses is greater in the set of hypotheses that are selected for prediction than in the set of all candidate hypotheses. This is captured in the following proposition.

Proposition 3. Let hypotheses selected via *prediction* and *HARKing* exhibit base rates π_p and π, respectively. Then prediction is more reliable than HARKing with respect to false discovery, $FDR(M^p) < FDR(M^h)$, just in case $\pi_p > \pi$.

This explains why misrepresentation of hypotheses selected ex post as those selected ex ante to observing data can be pernicious. If we expect a researcher to have meaningful domain knowledge, then we should expect her choice of hypothesis to be informative and so for the hypothesis she selects ex ante to be more likely to be true. If we are misled on this count, and the hypothesis reported does not boast such support, then we will (intuitively) assign too great a credence to her findings.

Proposition 3 also tells us when HARKing will produce more reliable findings than prediction. Consider an unlucky world in which the scientist's judgment is anti-correlated with the truth. That is, when the researcher makes predictions, she selects false hypotheses at a rate that is worse than chance. In this world, it is better to HARK than to predict.

Such an unfortunate world is not just a modeling fantasy. It is well-documented that in the domains of social and political punditry, humans can perform the impressive feat of doing consistently worse than chance.[21] More generally, prediction may fare unfavorably when we are confronted with problems where our domain knowledge is limited—as in cases with limited theory, few or no prior studies, or where common sense is largely unhelpful. One can think of analyses of any complex system where it is to be expected that a multitude of factors conspire to produce effects of interest. In such cases, a small increase in the false discoveries produced by ex post hypothesis selection may be compensated for by a greater decrease in false omission rates that may redound to leads for future, confirmatory research.

In sum, prediction can yield greater false discovery rates than HARKing, or HARKing can produce greater false discovery; what determines which obtains is the prior odds of hypotheses submitted to each. In the real world, we can expect that the prior odds of hypotheses submitted for HARKing is determined by the challenge of the domain, and the difference of prior odds of hypotheses submitted for prediction is determined by researcher judgment.

[21] See Tetlock (2017) for an excellent presentation of the literature on expert political judgment.

1.6 Jumping the HARK: From Misdiagnosis to Misprescription

Misdiagnosis of HARKing can ramify in the misprescription of solutions to the replication crisis. One prominent line of thinking in the literature is that if questionable research practices such as HARKing are bad because they make studies more likely to yield false positive results, then one natural solution is to lower the conventional threshold for statistical significance to compensate. A recent statement signed by over 50 prominent methodologists proposes redefining statistical significance in just this way (Benjamin et al., 2018). Stated plainly,

> "For fields where the threshold for defining statistical significance for new discoveries is $p < 0.05$, we propose a change to $p < 0.005$. This simple step would immediately improve the reproducibility of scientific research in many fields." (p. 6)

This is not a prima facie unreasonable proposal. Fields such as genomics and high energy physics have profitably set more stringent standards for their conventional significance threshold in the context of particle detection and gene-wide association studies.[22] But there are differences in methods between fields that can make a difference here. John Ioannidis expresses worry regarding the efficacy of lowering the significance threshold in the social and biomedical sciences, citing the relatively greater researcher degrees of freedom in those disciplines, "Adopting lower p-value thresholds may... [produce] collateral harms... bias may escalate rather than decrease if researchers... try to find ways to make the results have lower p-values" (Ioannidis, 2018, p. 1430). The model we present can be seen as demonstrating a precise realization of Ioannidis' worry.

To see why such an intervention can been seen as a solution to the ills of questionable research practices such as HARKing, consider the following simple model where researchers engage in a strategic mixture of both prediction and HARKing protocols: she follows a protocol of prediction when she can, and a protocol of HARKing when she must in order to attain statistical significance for some findings, and so to publish her study. Call this method *fallback Harking* M^{fh}. This can be thought of as a plausible approximation of what many researchers in certain domains in fact do (John et al., 2012).

Here, a study consists of procuring data against which a set of N logically independent hypotheses $\{H_i\}_{i=1}^{N}$ may be tested. In fallback HARKing, prior

[22] For a recent assessment of significance thresholds in GWAS see Panagiotou et al. (2012). For the history of the five-σ rule see Franklin (2013). Though, for a critical assessment of the latter see Lyons (2013) and (2015).

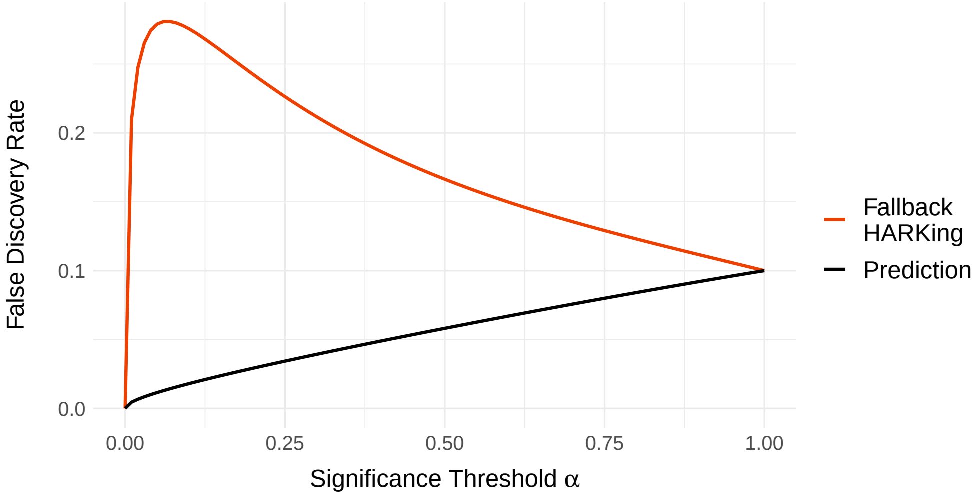


Figure 1.2: The false discovery rates for fallback HARKing and prediction protocols as functions of the significance threshold α when $\pi_p = 0.9, \pi = 0.1$, and $\beta = 0.2$.

to observing the data, a researcher selects the hypothesis for testing, H_1, that she judges is most likely true. Upon observing the data, if she finds that her hypothesis, H_1, is statistically significant then she reports it. If, on the other hand, she finds that her hypothesis is not statistically significant, she casts a broader net and turns to the $N-1$ other possible hypotheses $\{H_i\}_{i=2}^{N}$ and reports one that is significant, if such a one exists.[23]

Importantly, the researcher's prediction here is informed by her domain knowledge. The hypothesis she chooses prior to observing the data, H_1, is supported by some combination of theory, previous results, common sense, and so on, and so the hypotheses she chooses are on average more likely to be true than a random member of the other $N-1$ hypotheses, $P(H_1) = \pi_1 > \mathbb{E}[\{\pi_i\}_{i=2}^{N}]$. These other hypothesis may be true, but they are not, on average, as well supported by her domain knowledge.

Consider the effects of lowering the conventional threshold for statistical

[23] As with HARKing, a hypothesis is chosen at random from among the set of significant hypotheses. One could instead consider the case where the researcher chooses the hypothesis with the lowest p-value from among the set of statistically significant hypotheses. The qualitative outcome of the model—the possibility of an increase of false discovery rate as the significance threshold is lowered—would not change so long as researcher judgment was sufficiently informed.

significance, α, in such a case—where researchers are engaged in a reporting protocol like fallback HARKing. Figure 1.2 shows the relationship between the significance threshold, α, and false discovery, FDR, in such a case. If all researchers adhere to the method of prediction, M^p, then lowering α expectably reduces the false discovery rate $FDR(M^p(\alpha))$ (see the black line in Figure 1.2). Similarly, if all researchers adhere to a protocol of HARKing, then false discovery rate decreases as α decreases.[24]

However, if researchers follow a protocol like fallback HARKing, $M^{fb}(\alpha)$, the false discovery rate can actually increase (see the red line in Figure 1.2). The reason for this is that, as α decreases—and the evidential standard for significance becomes more stringent—researchers are less likely to attain statistical significance for their primary hypotheses, and so they are more likely to turn to scouring auxiliary hypotheses. That is, researchers must turn away from the few hypotheses with greater prior odds and toward the many hypotheses that exhibit potentially far lower prior odds. And since there are many more of the latter than the former, they are likely to find some that have attained significance by chance. These statistically significant 'fallback hypotheses', in turn, are more likely to be false discoveries.[25]

In short, when researcher judgment on a problem is well-informed, lowering the significance threshold can actually push researchers off of their fewer more promising hypotheses and onto dredging their many unpromising ones, and so can increase the false discovery rate of a population of studies.

When should we expect that lowering the significance threshold will increase false discovery? This is an empirical question with deep implications; and the answer will depend on several key factors. In particular, it will depend on the prevalence of behaviors approximating fallback HARKing,[26] the current value of the significance threshold and the extent to which it is lowered, the average power of the population of studies under question,[27] the prevalence of true hypotheses in the domain in question,[28] and, as our analysis reveals, the strength of researcher judgment.[29] Exploration of optimal values for statistical

[24] Though, of course, by the same token, lowering α must increase the false omission rate—a greater fraction of cases where the alternative hypothesis is true must fail to attain significance.

[25] See the Computational Appendix for links to the GUI and source code of a model of the statistical properties of prediction, HARKing, and fallback HARKing.

[26] For studies of the prevalence of questionable research practices, including HARKing, see John et al. (2012) and Head et al. (2015).

[27] See Szucs and Ioannidis (2017) and Lamberink et al. (2018) for estimates of the average power of studies in psychology and cognitive neuroscience, and clinical trials.

[28] For recent work on this count in replication markets, see Pawel and Held (2020).

[29] There is little work isolating the reliability of researcher judgment in hypothesis selection. However, recent work on replication markets shows that researchers predict the replication of

significance for a given domain given the distinctive methods and challenges of that domain must remain for future work.

Note, however, that correctly identifying the mechanism by which the HARKing affects the reliability of studies—differences in expected prior odds of hypotheses selected ex ante and ex post observing data—was crucial for identifying the *very possibility* of the undesirable consequences of redefining statistical significance just discussed. The effect could not be characterized by merely looking at either the Type I or II error rates of protocols, or by examining their p-values or test severity. One must attend to the reporting protocols with the prior odds of hypotheses each characteristically submits to tests.

1.7 HARKening Back to the Good Old Bayes'

We have seen that the standard accounts in the literature fail to explain why, precisely, HARKing undermines the reliability of scientific findings. The properties they claim account for why HARKing is bad obtain even when HARKing improves the reliability of findings. A Bayesian analysis elucidates the relationship between HARKing and the reliability of scientific findings: HARKing can increase or decrease the reliability of findings relative to prediction as a function of the differences in the prior odds of hypotheses characteristically selected ex ante and ex post to observing data.

Further, we have conjectured that the natural mechanism for producing the difference in prior odds of hypotheses selected in prediction is researcher judgment. When a scientist is meaningfully informed in her ex ante choice of hypothesis then her prediction is formally equivalent to restricting the set of reported study hypotheses to a subset which will, on average, be more likely to be true. HARKing, on the other hand, is formally equivalent to failing to make such a restriction. Thus, when scientists are uninformed, or worse, systematically biased, prediction can correspond to restriction to a subset of hypotheses with lower expected prior odds, and so HARKing, or a plausible variant of HARKing, can outperform prediction in terms of the false discovery and omission rates of the findings produced.[30]

the studies of their colleagues with remarkable accuracy (Dreber et al., 2015). Though, of course, it is possible that this good judgment extends to the hypotheses of others, and not to one's own. Further, such findings may produce adequate estimates for the replicability of studies, if not the pre-test prevalence of true hypotheses; though this suggests natural methods to estimate the latter using the former.

[30] This provides a recommendation for when selecting hypotheses ex post is unequivocally likely to be better: in inference tasks where the set of plausible hypotheses tends to be very large, highly complicated, or where researchers are known to be biased.

Moreover, it is important to bear in mind that the decrease in the false discovery rate produced by prediction comes at the cost of increase in false omission rates. Any non-trivial restriction of the set of candidate hypotheses must lower the absolute rate of discovery. This suggests that a more ethical version of HARKing, such as transparent HARKing, may be preferable in contexts where lowering the false omission rate matters more to researchers or policymakers than lowering false discovery rates.[31]

We have also argued that the misunderstanding of HARKing has consequences for proposed solutions to the replication crisis. In particular, we considered a recent proposal to redefine the conventional threshold for statistical significance and how such a proposal can lead to undesirable consequences in light of an accurate understanding of how HARKing affects the reliability of findings.

Our moral is that current accounts of HARKing that stem from a frequentist philosophy of statistics fail to explain the actual logic of its interaction with the reliability of scientific findings, that their misdiagnosis ramifies into misprescription for solutions in the context of the replication crisis, and that a Bayesian analysis of the problem makes this all clear.

1.8 Mathematical Appendix

For the following proofs it is assumed that we have non-extremal values for each the significance threshold, $\alpha \in (0, 1)$, and mean power, $1 - \beta \in (0, 1)$, of tests as well as for the prevalence of true hypothesis, $\pi \in (0, 1)$, in the set of possible hypotheses.

Proof of Proposition 1. Consider the false discovery rate for the prediction protocol M^p. Recall that, in prediction, a hypothesis is selected for testing prior to observing the data, and that after observing the data the selected hypothesis is reported only if it attains statistical significance.

Let $\pi = Pr(H_i)$ be the fraction of true hypotheses in the set of possible hypotheses; $\pi_p = Pr(H_1)$ the fraction of true hypotheses in the subset of hypotheses selected via prediction (first we consider the case where a hypothesis

[31] One can think of the characteristic differences in such preferences in the concrete examples of preliminary, exploratory analyses aimed at identifying promising future cancer treatments in contrast to confirmatory analyses of phase III clinical trials aimed at vetting whether a pharmaceutical should hit the shelves and be made available to the public.

is selected at random for prediction; and so $\pi_p = \pi$); let $p = Pr(T > t|H_0)$ denote the p-value of the test; $\alpha = Pr(p \leq \alpha|H_0)$ the significance threshold of the test; and $1 - \beta = Pr(p \leq \alpha|H_1)$ the power of the test. The false discovery rate for prediction is then obtained via Bayes rule.

$$\begin{aligned} FDR(M^h) &= \frac{Pr(p \leq \alpha, H_0)}{Pr(p \leq \alpha)} \\ &= \frac{Pr(p \leq \alpha|H_0)Pr(H_0)}{Pr(p \leq \alpha|H_0)Pr(H_0) + Pr(p \leq \alpha|H_1)Pr(H_1)} \\ &= \frac{\alpha(1-\pi_p)}{\alpha(1-\pi_p) + (1-\beta)\pi_p} \\ &= \Big(1 + \frac{\pi_p}{1-\pi_p}\frac{1-\beta}{\alpha}\Big)^{-1}. \end{aligned} \tag{A}$$

Next, consider the false discovery rate for the HARKing protocol M^h. Recall that, in HARKing, a hypothesis is selected for reporting at random after observing the data from the set of hypotheses that are significant (if the set is nonempty).

Let π, α and β be as before, let p_ℓ denote the p-value of hypothesis with index ℓ, and let z_ℓ be a random variable such that $z_\ell = 1$ if hypothesis ℓ is selected to be reported and $z_\ell = 0$ otherwise. We obtain the false discovery rate for a hypothesis H^ℓ selected via HARKing as follows. Note that, by stipulation, if H^ℓ was selected for reporting, then it was in the set of significant hypotheses. We have

$$FDR(M^h) = \frac{Pr(p_\ell \leq \alpha, H_0^\ell, z_\ell = 1)}{Pr(p_\ell \leq \alpha, z_\ell = 1)}.$$

The likelihood-prior products can be expanded to $Pr(z_\ell = 1|p_\ell \leq \alpha, H_j^\ell)Pr(p_\ell \leq \alpha|H_j^\ell)Pr(H_0^\ell)$ for $j = 0, 1$. Note that the probability that H^ℓ is significant is independent of its truth conditional on its p-value. Hence, we have $Pr(z_\ell = 1|p_\ell \leq \alpha, H_j^\ell) = Pr(z_\ell = 1|p_\ell \leq \alpha)$, which cancels out in the numerator and (expanded) denominator. This leaves

$$\begin{aligned} &= \frac{Pr(p_\ell \leq \alpha|H_0^\ell)Pr(H_0^\ell)}{Pr(p_\ell \leq \alpha|H_0^\ell)Pr(H_0^\ell) + Pr(p_\ell \leq \alpha|H_1^\ell)Pr(H_1^\ell)} \\ &= \frac{\alpha(1-\pi)}{\alpha(1-\pi) + (1-\beta)\pi} \\ &= \Big(1 + \frac{\pi}{1-\pi}\frac{1-\beta}{\alpha}\Big)^{-1}. \end{aligned} \tag{B}$$

Now, compare equations (A) and (B), capturing the false discovery rates of prediction and HARKing protocols, respectively. When the fraction of true hypotheses selected for testing under prediction is the same as the fraction of true possible hypotheses, $\pi_p = \pi$, equations (A) and (B) are equal, and hence $FDR(M^p) = FDR(M^h)$, as desired. □

Proof of Proposition 2. To consider the false discovery rate of selective HARKing M^{sh}, let ℓ denote the the statistically significant hypothesis with the lowest p-value selected from the set of L hypotheses. That is, ℓ is the index of the hypothesis reported by selective HARKing. Let $-\ell'$ be the index of the hypothesis with the next-lowest p-value $p^*_{-\ell} = \inf_{\ell' \neq \ell} p_{\ell'}$.

The false discovery rate of selective HARKing then is just equal to the expectation of the false discovery rate of the hypothesis reported by selective HARKing, $FDR(M^{sh}) = \mathbb{E}_\ell[FDR(M^{sh}_\ell)]$. Further, the false discovery rate of the selected hypothesis is equal to its expectation under the distribution of the p-values of the hypothesis with the next-lowest p-value $FDR(M^{sh}_\ell) = \mathbb{E}_{p^*_{-\ell}}[FDR(M^{sh}_\ell(p^*_{-\ell}))]$. Deriving the false discovery rate of $M^{sh}_\ell(p^*_{-\ell})$ just as before we get

$$\begin{aligned} FDR(M^{sh}_\ell(p^*_{-\ell})) &= \Big(1 + \frac{\pi}{1-\pi}\frac{1-\beta(\min\{\alpha, p^*_\ell\})}{\min\{\alpha, p^*_\ell\}}\Big)^{-1} \\ &\leq \Big(1 + \frac{\pi}{1-\pi}\frac{1-\beta}{\alpha}\Big)^{-1} = FDR(M^p) \end{aligned}$$

Thus $FDR(M^{sh}) \leq FDR(M^h) = FDR(M^p)$ as desired.

To prove the false discovery rate for selective HARKing is decreasing in the number of hypotheses, first note that $P^{sh}_\ell(p^*_{-\ell})$ is increasing in $p^*_{-\ell}$ for $p^*_{-\ell} < \alpha$ and constant otherwise. The CDF of $p^*_{-\ell}$ given the number of hypotheses L is $Pr(p^*_{-\ell} \leq t|L) = Pr(p^*_{-\ell} \leq t)^{L-1}$ and thus $Pr(p^*_{-\ell} \leq t|L)$ is first-order stochastically dominated by $Pr(p^*_{-\ell} \leq t|L')$ for $L' < L$. It follows that P^{sh} is decreasing in L. □

Proof of Proposition 3. Finally, we show that prediction is more reliable than HARKing, $FDR(M^p) > FDR(M^h)$ just in case $\pi_p > \pi$. Consider a population of studies and let π be the fraction of true hypotheses in the set of possible hypotheses; let π_p be the fraction of true hypotheses in the subset of hypotheses selected via prediction; α the significance threshold of tests; and $1-\beta$ their power.

From the preceding proofs, we have the following false discovery rates for the prediction and HARKing protocols:

$$FDR(M^p) = \Big(1 + \frac{\pi_p}{1-\pi_p}\frac{1-\beta}{\alpha}\Big)^{-1}, \quad \text{and} \quad FDR(M^h) = \Big(1 + \frac{\pi}{1-\pi}\frac{1-\beta}{\alpha}\Big)^{-1}.$$

So $FDR(M^p)$ and $FDR(M^h)$, as established in proposition 1.2, are equal when the fraction of true hypotheses selected under either protocol are equal, $\pi_p = \pi$. Thus, all that remains to be shown is that the false discovery rate of prediction is decreasing in π_p. For this, we simply take the derivative of $FDR(M^p)$ with respect to π_p and show that it is negative.

$$\frac{\partial}{\partial \pi_p}[FDR(M^p)] = \frac{\alpha(\beta-1)}{(\alpha(\pi_p-1)+(\beta-1)\pi_p)^2} < 0.$$

And the expression is negative since the numerator is negative for the assumed values of type I and II error rates ($\alpha, \beta \in (0,1)$) and since the denominator must be positive. □

Proposition 4. *When hypotheses are selected from the same set of candidate hypotheses with fraction $\pi_p = \pi \in (0,1)$ true hypotheses, then prediction yields the same false omission rate as HARKing.*

Proof. Let π_p, π, p, α, and β be as before. Consider the false omission rate of prediction, M^p.

$$\begin{aligned} FOR(M^p) &= \frac{Pr(p > \alpha, H_1)}{Pr(p > \alpha)} \\ &= \frac{Pr(p > \alpha | H_1) Pr(H_1)}{Pr(p > \alpha | H_1) Pr(H_1) + Pr(p > \alpha | H_0) Pr(H_0)} \\ &= \frac{\beta \pi_p}{\beta \pi_p + (1-\alpha)(1-\pi_p)} \\ &= \Big(1 + \frac{1-\pi_p}{\pi_p}\frac{1-\alpha}{\beta}\Big)^{-1}. \end{aligned} \tag{A}$$

Next, consider the false omission rate of the HARKing protocol, M^h.

$$\begin{aligned} FDR(M^h) &= \frac{Pr(p_\ell \leq \alpha, H_0^\ell, z_\ell = 1)}{Pr(p_\ell \leq \alpha, z_\ell = 1)}. \\ &= \frac{Pr(p_\ell \leq \alpha | H_0^\ell) Pr(H_0^\ell)}{Pr(p_\ell \leq \alpha | H_0^\ell) Pr(H_0^\ell) + Pr(p_\ell \leq \alpha | H_1^\ell) Pr(H_1^\ell)} \end{aligned}$$

$$= \frac{\alpha(1-\pi)}{\alpha(1-\pi)+(1-\beta)\pi}$$
$$= \Big(1+\frac{1-\pi}{\pi}\frac{1-\alpha}{\beta}\Big)^{-1}. \tag{B}$$

Clearly, whenever $\pi_p = \pi$, the false omission rates given in (A) and (B) are equal, as desired. □

Proposition 5. *Prediction is more reliable than HARKing with respect to false omission, just in case $\pi_p < \pi$.*

Proof. Let π_p, π, p, α, and β be as before. Consider the false omission rate of prediction, M^p. From the preceding proofs, we have the following false omission rates for the prediction and HARKing protocols:

$$FOR(M^p) = \Big(1+\frac{1-\pi_p}{\pi_p}\frac{1-\alpha}{\beta}\Big)^{-1}, \quad \text{and} \quad FOR(M^h) = \Big(1+\frac{1-\pi}{\pi}\frac{1-\alpha}{\beta}\Big)^{-1}.$$

Now, observe that false omission rate of prediction is increasing in π_p. For this, take the derivative of $F0R(M^p)$ with respect to π_p and show that it is positive.

$$\frac{\partial}{\partial \pi_p}[FOR(M^p)] = \frac{\beta(1-\alpha)}{(\alpha(\pi_p-1)+(\beta-1)\pi_p+1)^2} > 0,$$

Since the numerator is positive (given $0 < \alpha, \beta < 1$) as is the denominator. □

1.9 Computational Appendix

A GUI for exploring the performance of reporting protocols for prediction, HARKing, and fallback HARKing is available at: `https://amohseni.shinyapps.io/Reporting-Protocols-and-the-Reliability-of-Science/`.

The R code for the computational model can be found at: `https://github.com/amohseni/Reporting-Protocols-and-the-Reliability-of-Science/`

Truth and Conformity on Networks 2

Abstract

Typically, public discussions of questions of social import exhibit two important properties: (1) they are influenced by conformity bias, and (2) the influence of conformity is expressed via social networks. We examine how social learning on networks proceeds under the influence of conformity bias. In our model, heterogeneous agents express public opinions where those expressions are driven by the competing priorities of accuracy and of conformity to one's peers. Agents learn, by Bayesian conditionalization, from private evidence from nature, and from the public declarations of other agents. Our key findings are that networks that produce configurations of social relationships that sustain a diversity of opinions empower honest communication and reliable acquisition of true beliefs, and that the networks that do this best turn out to be those which are both less centralized and less connected.

2.1 Introduction

Epistemology is the study of true, justified, or reliable beliefs. Social epistemology is the study of the effect of social structures and interactions on the emergence and maintenance of such beliefs. Much of the recent work in social epistemology has been in the application of game-theoretic techniques to modeling communities of rational inquirers—epistemic communities—to see what incentive and interaction structures are conducive to the outcomes of accuracy,[1] efficiency,[2] and equity.[3]

One of the most important domains of social inquiry is that of broad public discourse. Which social policy will lead to better outcomes? Which political candidate is more qualified for office? Typically, public discussion on such questions of import is influenced by the human tendency of conformity. Individual decisions are informed and influenced by peers; the presence of conformist bias in social discourse is well-studied, and well-supported.[4]

We present a model of social inquiry where it exhibits two properties endemic to matters of public discussion: (1) individuals are subject to varying degrees to conformity bias, and (2) the influence of the pressure to conformity is expressed via social networks. We examine how the structure of social ties in tandem with conformity bias can influence the flow and reliability of information in matters of public opinion.

In our model, heterogeneous agents express public opinions where those expressions are driven by the competing priorities of accuracy and of conformity. Agents learn, by Bayesian conditionalization, from private evidence from nature, and from the public declarations of other agents.

Several key findings emerge. We see that the most influential public declarations are made by agents when they go against the consensus of their neighbors, but that the most informative declarations, on average, are made by agents when their social influences are balanced. This provides a unifying explanation for our results: networks that produce configurations of social relationships that sustain a diversity of opinions empower honest communication and hence reliable acquisition of the truth.

In related literature on network epistemology (Zollman, 2007, 2010, 2013), less connected networks are shown, under the right conditions, to increase the reliability of inquiry. In those cases, greater connectivity can cause premature "lock-in" to consensus in epistemic communities dealing with an exploration-exploitation trade-off.[5] We arrive at a similar moral by different means.

We show that networks are differentially conducive to informative communication depending on the degree to which a community is divided in its publicly stated opinions. When communities are most divided, more connected networks, such as complete networks, do best. Whereas, when communities are near

[1] See Zollman (2007, 2009, 2013), Mayo-Wilson et al. (2013), and Grim et al. (2013).

[2] See Heesen (2017), Kitcher (1990, 1993), and Strevens (2003, 2013).

[3] See O'Connor et al., and Bruner and O'Connor (2016)

[4] See Asch (1955), Bond and Smith (1996), and Morganand and Laland (2012).

[5] See (Rosenstock et al., 2017) for an analysis of the specific conditions under which the effect described in (Zollman, 2007) obtains.

consensus, less connected networks exhibiting low degree-centrality, such as circle networks, are optimal.

Across the networks literature, star networks have been shown to possess certain optimality properties: they emerge as the product of various processes of strategic network formation (Goeree et al., 2009; Barrett et al., 2017), can lead to efficient division of cognitive labor (Goyal, 2007; Zollman, 2013), and can provide optimal conditions for information dissemination (Goeree et al., 2009). In contrast, we find that, in the presence of a modicum of conformity bias, star networks produce to the worst possible conditions for social learning.

Our analysis has implications for broad concerns in social epistemology. For example, arguments for the merits of deliberative democracy turn on the relative success of epistemic communities in engaging collective inquiry, discourse, and decision (Landemore, 2012; Mercier and Landemore, 2012). Our results provide a distinct justification for the import of a diversity of opinions in such contexts. And, if collective intelligence is to justify democratic institutions and practices, then it behooves us to identify and promote (or resist) the social structures which conduce to (or derange) the reliability of public discourse and so the prospects of a flourishing deliberative democracy.

In §2.2, we explain our model. In §2.3, we present the long run success of learning in the presence of conformist bias. In §2.4, we present simulations illustrating our central results. In §2.5, we provide an analysis of the deeper patterns that unify and explain our results. In §2.6, we conclude.

2.2 The Model

A Vignette: Caesar or Pompeia?

To animate our model, let us consider an anachronistic allegorical vignette. A community of Roman citizens has come together to discuss which candidate is better qualified for office. The candidates are Caesar and Pompeia. In discussing their beliefs, the citizens are influenced, to varying degrees, by two competing motivations: the motivation to say, honestly, who they believe is the better qualified candidate, and the motivation to agree with their neighbors, or, more particularly, those with whom they share social or economic ties.

Each citizen varies with respect to the weight she places on each honesty and conformity. On one extreme, we may find Titus the Truth-Teller, who speaks his mind, come what may. Titus has come to believe—both from what is public knowledge, and from his own private information and experiences—that it is Pompeia who is more likely to be a better candidate. And so he declares as much, and he does so without any thought or worry concerning the impact of his declaration on the social regard of his peers.

On the other extreme, we find Cassius the Conformist, for whom harmony with peers is his sole concern. Making his true beliefs known does not enter the picture. Whichever candidate his peers favor, Cassius favors. Now, it happens to be Caesar.

Most of the remaining citizens, however, are not so extreme in their dispositions, but rather fall somewhere between Titus and Cassius. They care about making their true beliefs known, to some degree, and also about harmony with their peers, to some degree. Most make their declarations in a way that is contingent both on the strength of their beliefs, and on the weight of the social pressures around them.

In such a community, individuals come to private beliefs about which candidate is more qualified using their private evidence as well as what they can glean from the public declarations of others. Each individual makes her public declaration in turn–one that reflects her best interests and so is informed by her current beliefs, by the particular social pressures she experiences, and by the degree to which she is motivated by each. This process is repeated, and so the private beliefs and public declarations of the community evolve over time. Our model provides a general formulation of how such learning and discourse may unfold and how it is influenced by the structure of social and economic ties that underpin the community.

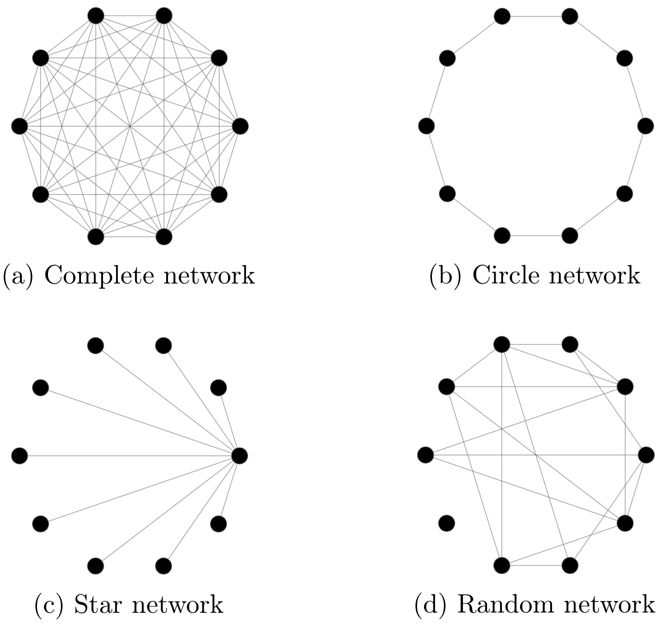
Formal Description of the Model

Imagine that there are two states of the world: θ and $\neg\theta$. We can think of these as corresponding to where one social policy will lead to better outcomes, or one political candidate will be better suited to office.

Agents are interested in learning the true state of the world. This proceeds in two ways: (1) They get private evidence $\sigma \sim f_\theta(\sigma)$ from Nature; we can think of these as hearing some piece of news, or reasoning about an argument. And (2), they observe the public declarations $\mathbf{x}_{-i} \in \{\theta, \neg\theta\}^{N-1}$ of other agents across the network. Declarations indicate to others the state a declaring agent ostensibly believes to be true.

For each agent i, her payoffs are a convex combination of her *truth-seeking* orientation α_i and desire for *conformity* to her neighbors $(1-\alpha_i)$. Her payoff for a declaration $x \in \{\theta, \neg\theta\}$ then is given by

$$U_i(x) = \alpha_i P_i(x) + (1-\alpha_i) N_i(x)$$



(a) Complete network

(b) Circle network

(c) Star network

(d) Random network

Figure 2.1: Social networks with 10 agents.

where $P_i(x)$ is agent i's expectation of the truth of x given her current information, and $N_i(x)$ is the proportion of her neighbors that have also declared x.[6]

We can think of an agent i as engaged in two games simultaneously which determine her payoffs in proportion to her type: a Bayesian learning game that contributes α_i of her payoff, where the data are the agent's private evidence σ and others' public declarations $\mathbf{x}_{-i}$, and an n_i-player pure coordination game that constitutes the remaining $(1 - \alpha_i)$ of her payoff, where n_i is the count of agent i's neighbors.

Our epistemic community of N agents inhabits a society where their patterns of shared social influence are described by a network.[7] Here, nodes represent agents, and neighbors are connected by edges. Standard networks include

[6] Note that an agent's truth-seeking payoff for a declaration is based on her expectation that it corresponds to the true state of the world—agents do not know, and do not find out, whether their assessments are accurate.

[7] In past work on social networks, the network has been used to represent each patterns of transmission of social influence and patterns of transmission of information. Here, we focus on the effect of patterns of social influence, and so the network structure captures the former but not the latter, and we follow (Banerjee, 1992; Bikhchandani et al., 1992; Smith and Sørensen, 2000) in assuming that individual actions are observable to all individuals in the community.

complete, circle, star, and random networks (see Figure 2.1).

Networks vary with respect to the patterns of social influence they capture. The complete network describes a social structure in which each agent has social ties with every other. The circle describes a social structure in which each agent shares social ties with exactly two other individuals. Note that complete and circle networks are special cases of regular networks,[8] where the regular network is of degree $N-1$ and degree 2, respectively. In contrast, the star network describes a centralized social structure, where one individual (a central agent) has far more connections than the rest (the peripheral agents), who are otherwise socially isolated.

Before the game, agent types (truth-seeking/conformity orientations) are drawn from a continuous distribution:

$$\alpha_1, \ldots, \alpha_N \overset{iid}{\sim} G \text{ with } supp(G) = [0,1]$$

And Nature randomly chooses the state of the world to be θ or $\neg\theta$. Each state of the world induces a distinct distribution from which evidence $\sigma \sim f_\theta(\sigma)$ may be drawn.

The distributions $f_\theta(\sigma) = 2\sigma$ and $f_{\neg\theta}(\sigma) = 2 - 2\sigma$, depicted in Figure 2.2, are used in our simulations due to their convenient functional form. More generally, however, the distributions need only satisfy: *mutual absolute continuity* and *unbounded evidence*. Mutual absolute continuity requires that both distributions agree on what subsets of possible evidence have positive probability, meaning that no single piece of evidence can falsify one or the other. And unbounded evidence give us that evidence has the potential, in principle, to make one arbitrarily (though not completely) confident of either state. We take this to be a reasonable assumption, as we want to allow that, for any degree of belief shy of absolute certainty, there can—in principle—exist some evidence, however unlikely, which is sufficiently compelling to produce that belief.

In each round, an agent is chosen at random to receive private evidence from Nature, and to make a public declaration to be observed by the network. Upon receiving her evidence, an agent updates her beliefs, via conditionalization, about the true state of the world. This is done in the normal way, using Bayes' rule

$$P(\theta|\sigma, \boldsymbol{h}^t) = \frac{P(\sigma|\theta)P(\theta|\boldsymbol{h}^t)}{P(\sigma|\theta)P(\theta|\boldsymbol{h}^t) + P(\sigma|\neg\theta)P(\neg\theta|\boldsymbol{h}^t)}$$

where $P(\sigma|\theta)$ is the likelihood of her new evidence σ given the state θ, and $P(\theta|\boldsymbol{h}^t)$ is her prior on θ given the history of declarations at that time $\boldsymbol{h}^t$.

[8] Regular networks are those in which all nodes are of the same degree, or number of edges. Here, this will correspond to all agents having the same number of neighbors.

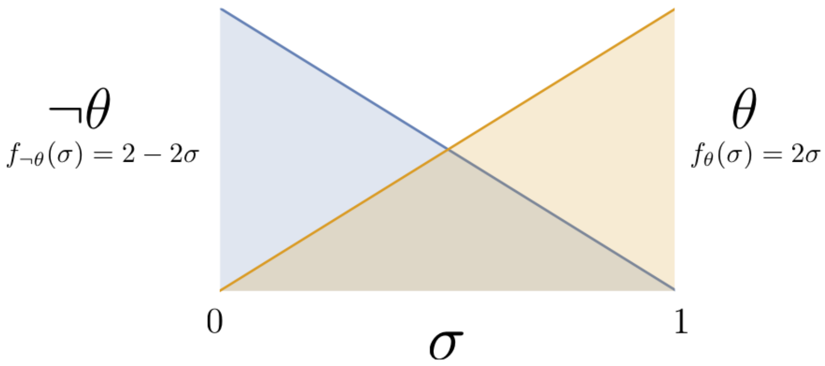


FIGURE 2.2: Distributions $f_\theta(\sigma)$, $f_{\neg\theta}(\sigma)$, of evidence σ, for each state of the world θ and $\neg\theta$.

Note that $P(\theta|\boldsymbol{h}^t)$ is also the public belief at that time—the shared portion of individual beliefs about the true state of the world constituted by the history of learning from public declarations. To simplify exposition, assume the population begins with ignorance priors.[9]

Next, the agent calculates her utilities, given her truth-seeking orientation, chooses her best response as a function of her private evidence and public prior (which, together, form her posterior probability $P(\theta|\sigma, \boldsymbol{h}^t)$ over θ), and the composition of her neighbors:[10]

$$BR_i(\sigma, N_i(x)) = \arg\max\{U_i(\theta), U_i(\neg\theta)\}.$$

Following this, the other agents in the network observe her declaration, and update their beliefs.[11] To do so, they must consider the likelihood of her having made her declaration given the composition of her neighbors, her likely evidence,

[9] An *ignorance prior* is a probability distribution assigning equal probability to all possibilities. Our proofs will require only non-degenerate priors, and our simulations will employ a range of priors.

[10] In the case of payoff ties, the agent chooses among her best responses at random.

[11] Note that individuals can observe the proportions of a declaring agent's neighbors making each declaration. We take this assumption to be plausible under some, but not all, conditions. In the context of public discourse, one can often observe–at least qualitatively–the social influences acting on other individuals. That is, when someone makes a declaration in favor of Caesar, we typically have a fair idea of whether her social network is predominantly pro-Caesar or pro-Pomepeia, some mixture of the two, and so on, and we use this information in assessing whether we think her assertion is more or less likely to be more or less socially or epistemically motivated. That said, future research exploring the effects of limiting observability of the network will be valuable.

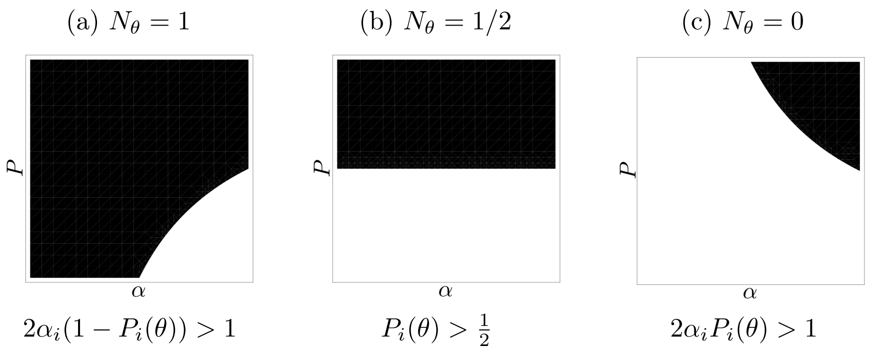


Figure 2.3: What is inferred from agent i's declaration of θ, as captured by condition (†), about her posterior belief $P_i(\theta)$ and truth-seeking orientation α_i, when different proportions of her neighbors $N_i(\theta) = 1, 1/2, 0$ are making the same declaration.

her possible truth-seeking orientations, and their own prior beliefs about the state of the world.

So, what precisely do agents learn from one another's declarations? Well, when agent i declares $x = \theta$, others know that it was her best response to do so. It follows that

$$\begin{aligned} U_i(\theta) &> U_i(\neg\theta) \\ \alpha_i P_i(\theta) + (1-\alpha)N_i(\theta) &> \alpha_i(1 - P_i(\theta)) + (1-\alpha_i)(1 - N_i(\theta)) \\ \alpha_i(2P_i(\theta) - 1) + (1-\alpha_i)(2N_i(\theta) - 1) &> 0. \end{aligned} \tag{†}$$

We can get an intuitive grasp of this inequality (†) by considering fixed values of the proportion of the declaring agent's neighbors who are also declaring θ (depicted in Figure 2.3.). The shaded area captures the values of agent i's truth-seeking orientation α_i (on the horizontal axis), and posterior belief P_i (on the vertical axis), that are compatible with her having declared θ. That is, the region in which (†) is satisfied.

Consider the $N_i(\theta) = 1$ case (Figure 2.3a). This is where all of the focal agent's neighbors are also declaring θ. Here, we see that a broad range of beliefs and truth-seeking orientations are compatible with her having declared θ. What can be ruled out (the area in white) is that it was *not* the case that she was both highly truth-seeking and strongly believed in the truth of θ. Here, others do not learn much from observing the focal agent's declaration.

Consider the $N_i(\theta) = 1/2$ case (Figure 2.3b). This is where the focal agent's neighbors are evenly split; half declaring θ and half $\neg\theta$. Here, the other agents infer that the focal agent's social influences are balanced, and so her truth-seeking orientation α_i is no longer relevant. Her declaration is now determined purely by her posterior belief. If $P_i(\theta) > 1/2$, then she would make the declaration she did, if not, she would not. Here, others learn the direction of the focal agent's belief, but not much about its strength.

Next, consider the $N_i(\theta) = 0$ case (FIGURE 2.3C). This is where none of the focal agent's neighbors are declaring θ. Here, we see that only a narrow range of beliefs and truth-seeking orientations are compatible with her declaration of θ. It must have been the case that she was both highly truth-seeking, and possess a strong belief in the truth of θ. Now, others learn both the direction and strength of the focal agent's belief, and through it about the strength of her evidence.

From such inferences, the agents in the population update their beliefs about the state of the world using Bayes' rule.[12] And so, in the ways described, rational agents learn from their own private evidence, the declarations of other agents in the network, and public belief about the true state evolves through discussion and across the network.

2.3 Truth in the Long Run

Our primary interest lies in dynamical analysis of the short-to-medium-run behavior of social inquiry under conformist bias. Before we proceed to this analysis, however, it may help us in this to understand the long-run trajectory of social learning under conformity. What we find is that, in the long run, irrespective of social structure or conformist bias, epistemic communities like the ones we have described will converge to believing in, and publicly declaring, the true state of the world.

More precisely, given any social network, unbounded evidence, and the possibility of sufficiently truth-seeking agents $1 \in \text{supp}(\alpha)$, a community of Bayesian learners will, with probability one, converge to knowing and declaring the truth in the long run. This is captured by the following proposition and its corollary.

Proposition 6. *An epistemic community learning about the state of the world will, in the long run, converge in belief to the true state.*[13]

Corollary 7. *For such an epistemic community, converging in belief to the true state implies converging to consensus in declaring the true state.*

[12] See APPENDIX A for the mathematical details.

[13] All proofs can be found in Appendix A.

Convergence in beliefs follows from the fact that our agents learn via Bayesian conditioning, that the true state is contained in each agent's hypothesis set, and that agent declarations are always to some degree informative as to the state of the world. Given this, classical convergence results for Bayesian learning[14] guarantee long run acquisition of the truth.

Convergence in declarations follows from the fact that, given convergence in beliefs, the community's beliefs will inevitably pass a threshold such that a consensus on declaring the true state cannot be escaped. Moreover, with enough time following the passing of this threshold of belief, the population will almost surely traverse a positive probability path to consensus on the true state, whereupon it will never leave this consensus.

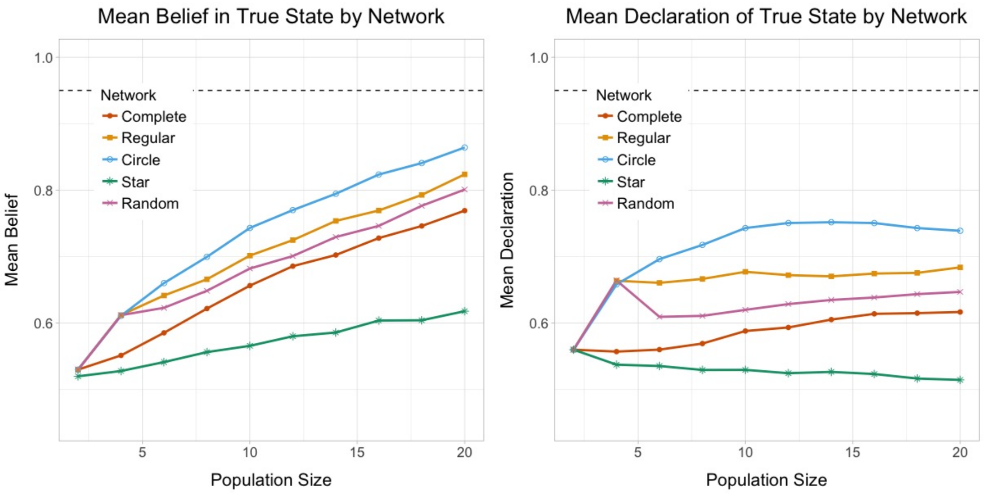
It may well be that "in the long run we are all dead," [Keynes, 1923, p. 80] but it can be helpful to confirm where we are headed. We have seen that our epistemic communities will arrive at the truth in the limit of time, so we turn to short and medium run analysis of social learning for a richer and more pressing picture of inquiry.

2.4 Truth and Conformity in the Short and Medium Run

What can be said about the short and medium run behavior of learning under conformity? What role does social structure play in the reliable acquisition of true beliefs? To answer these questions, we ran simulations of our model of epistemic communities engaged in social learning and discourse. We recorded and analyzed the resulting behavior over a parameter sweep of network types, population sizes, initial declarations, prior beliefs, and distributions of the individuals' truth-seeking and conformity orientations.

For the simulations, we varied the structure of social influences by placing our agents on each complete, regular (of degree $N/2$), circle, star, and random (of mean degree $N/2$) networks. We varied the number of agents N in the network from 2 agents (at which all networks are essentially identical) to 20 agents. We considered when the initial declarations of the society were at a consensus on the true state, a consensus on the false state, and an even split. We varied the shared prior beliefs of the population between relative confidence in the true state ($P(\theta) = 0.75$), skepticism toward the true state ($P(\theta) = 0.25$), and ambivalence about the true state ($P(\theta) = 0.5$). Each combination of network structure, population size, initial declarations, and prior beliefs composed one parameter setting.

[14] For an excellent exposition of the classic results, see Smith and Sørensen (2000).



(a)

(b)

FIGURE 2.4: Plots of the mean belief in the true state $P(\theta)$ (A), and declaration of the true state θ (B), for each network type, and for network sizes from 2 agents to 20. Note that the networks only become fully distinct at $N = 6$. The dashed line represents performance in total absence of any conformity.

For each parameter setting we ran 10,000 simulations where each simulation was composed of 100 turns, and where each turn consisted of the following phases: (1) a randomly selected agent receives her private evidence from Nature; (2) the agent updates her private belief in light of this evidence; (3) the agent chooses her best response given her beliefs, her neighbors' declarations, and her truth-seeking/conformity orientation; (4) the agent makes her declaration to the network; (5) the other agents in the network update their beliefs in light of her declaration.

Three regularities readily emerged from the data (see FIGURES 2.4A, 2.4B): (1) In all simulations, the star network performed worse than all other standard networks in terms of generating reliable belief in, and declaration of, the true state. (2) The circle network, on the other hand, performed better than other standard networks on all counts. (3) The other networks—complete, regular (of degree $N/2$), and random (of mean degree $N/2$)—yielded middling performances, neither as good as the circle, nor as poor as the star, with the regular network typically outperforming the random network, and the random network outperforming the complete network.[15]

[15] In our simulation plots (FIGURE 2.4), we mark the performance of learning in the absence

To make sense of these regularities in our simulation results, analytic treatment of the model and its dynamics is needed. What should be obvious is that conformity bias muddies the waters with respect to the information content of individuals' declarations. In the absence of conformity, our epistemic communities would rapidly and reliably acquire the truth, and the underlying network structure would make no difference to this learning.

What we will find is that different networks induce social configurations more or less conducive to honest communication, and that this will also depend on the degree to which the population is divided or unified in their public declarations.

2.5 Influence, Information, and Social Structure

To understand why different social networks are more or less conducive to the reliable acquisition of true beliefs, we first need a measure of informativeness. For this, we introduce the concepts of influence and informativeness of declarations, and show how they are related.

We define the *influence* of a declaration $x \in \{\theta, \neg\theta\}$ as the difference between the public belief in x before and after its declaration to the network, $q(x|x) - q(x)$, where q is the public belief. Next, we define the *informativeness* of a declaration $x \in \{\theta, \neg\theta\}$, as the reduction in uncertainty it produces with respect to its corresponding state when starting from a maximal entropy prior, $H(q|q(x) = 1/2) - H(q|x)$, where H is the Shannon entropy function.

We now derive the fact that the informativeness of a declaration is monotonically increasing in its influence on the public belief (see Lemma 14 in Appendix A). This gives us that a declaration will be (minimally) maximally influential just in case it is (minimally) maximally informative. We will use this fact repeatedly to infer the relative informativeness of declarations from their influence.

Optimal Information From Going Against the Grain

Given our measures of influence and informativeness, our first insight follows straightforwardly from our model of agents learning via Bayesian conditioning under uncertainty about one another's evidence and truth-seeking orientations. It is that the most informative declarations—those that have the most significant effect on the public belief—are those that "go against the grain." That is, those made by agents exactly when they deviate from the consensus of their neighbors.

of any conformity bias—that is, of unimpeded Bayesian learning—with a dashed line. We will continue to compare our results to this control case, denoting the case of learning in the absence of conformity bias in further plots (Figure 2.5, 2.6, 2.7) each time with a dashed line.

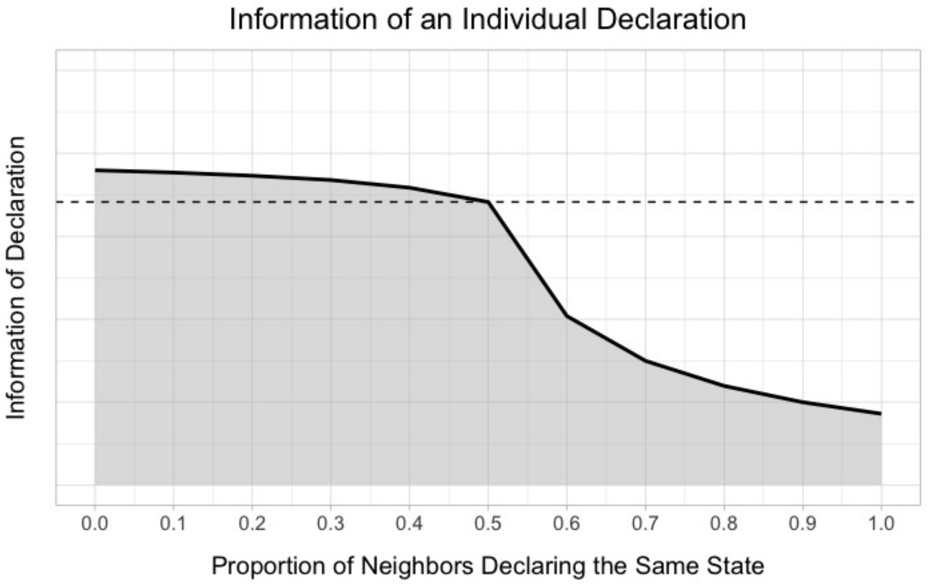


FIGURE 2.5: The influence and informativeness of an agent's declaration, as a function of the proportion of her neighbor's who are declaring the same state.

This insight is captured by the following proposition:

Proposition 8. *The informativeness of an agent's declaration is monotonically increasing in proportion of her neighbors who are declaring the opposing state.*

And since the minimum proportion of an agent's neighbors who may declare in favor of any state is zero, we have the following as an immediate corollary:

Corollary 9. *The most informative declaration in favor of a state is one made by an agent when she goes against the consensus of her neighbors.*

This corresponds to the case in FIGURE 2.3, where $N_i(\theta) = 0$, and is visualized by the plot of information of declarations in FIGURE 2.5 where we see the change in belief by the population in response to an agent's declaration as a function of the proportion of that agent's neighbors who are declaring the same state.

When an agent deviates from the consensus of her immediate peers, it is inferred by the broader network that she is both likely to be more truth-seeking and that she has received sufficiently strong evidence to justify the loss in social payoffs she incurred. No other declaration is more influential on the public belief.

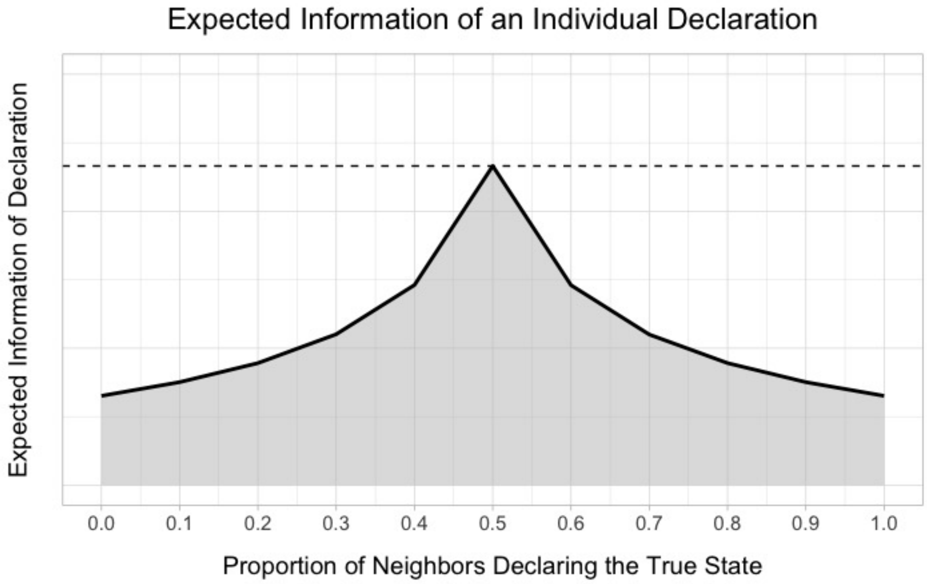


Figure 2.6: The *expected* influence and informativeness of an agent's declaration, as a function of the proportion of her neighbor's who are declaring the same state.

Optimal Expected Information From Conflicted Neighbors

We have seen that the most informative declarations occur when an agent goes against the consensus of her peers. But such declarations are rare, as it takes highly truth-seeking agents with good evidence to be willing to make them. We should ask then: under what conditions, *on average*, do we expect to find the most informative declarations?

These turns out to be the obverse of where we find the most influential declaration. The most informative declarations, on average, must come from individuals whose neighbors are perfectly divided in terms of their declarations.

This is captured by the following observations:

Observation 1. *The most influential and informative declaration, in expectation, is that made by an agent when her neighbors are evenly divided in their declarations.*[16]

[16] Our observations are computationally verified for the following distributions of types and evidence: the distribution of truth-seeking orientations in the population was varied from Beta(1,5) (corresponding to high conformism), to uniform, and Beta(5,1) (corresponding to high truth-seeking). And the distributions of evidence induced by each state of the world were

Observation 2. *The expected information of declarations is convex and increasing for $N_i(\theta) \in (0, 1/2)$ and convex and decreasing for $N_i(\theta) \in (1/2, 1)$.*

This corresponds to the case in Figure 2.3, where $N_i(\theta) = 1/2$, and is visualized by the plot of expected information of declarations in Figure 2.6. In Figure 2.6, we see the expected change in belief of the population in response to an agent's declaration, as a function of the proportion of that agent's neighbors who are declaring the true state. Our propositions make use of these observations.

It is when an agent's social influences equally represent each viable position that she is most free to declare her honest belief, and in such a case others infer that she is most likely doing so.

Informativeness of Networks

Which networks then are most conducive to the social configurations that yield honest communication? Using the insights developed so far, we extend the concept of expected informativeness to the level of social networks.

Assume that θ is the true state of the world, then *expected influence* of declarations $X = \{\theta, \neg\theta\}$ for an N-agent network $\mathcal{G}$ is given by

$$E_X[q(\theta|\mathcal{G}) - q(\theta)] \propto \sum_{k=0}^{N} \sum_{j=1}^{\binom{N}{k}} \sum_{i=1}^{N} E_X[q(\theta|x_i) - q(\theta)]$$

where the first sum is over the number of the agents in the network declaring the true state, the second sum is over the possible configurations of declarations in the network given the number of agents declaring the true state, and the third sum is over the individuals in the network.[17] In this way, we infer the informativeness of a network in aggregate as well as for fixed proportions of the community declaring the true state.

With a generalized measure of expected informativeness, we compute the expected informativeness of 10-agent networks for different proportions of the population declaring the true state (see Figure 2.7).

From this, several observations emerge. Denote the proportion of the community declaring θ by N_θ. For all networks, then, the least informative state is that of consensus, $N_\theta = 0$ or 1, and the most informative state is when

varied between the linear case described before, and Gaussian distributions with means of 1 and -1, and variances of 1, 10, and 100.

[17] Note that we have omitted the normalizing term from the definition of the influence of a declaration.

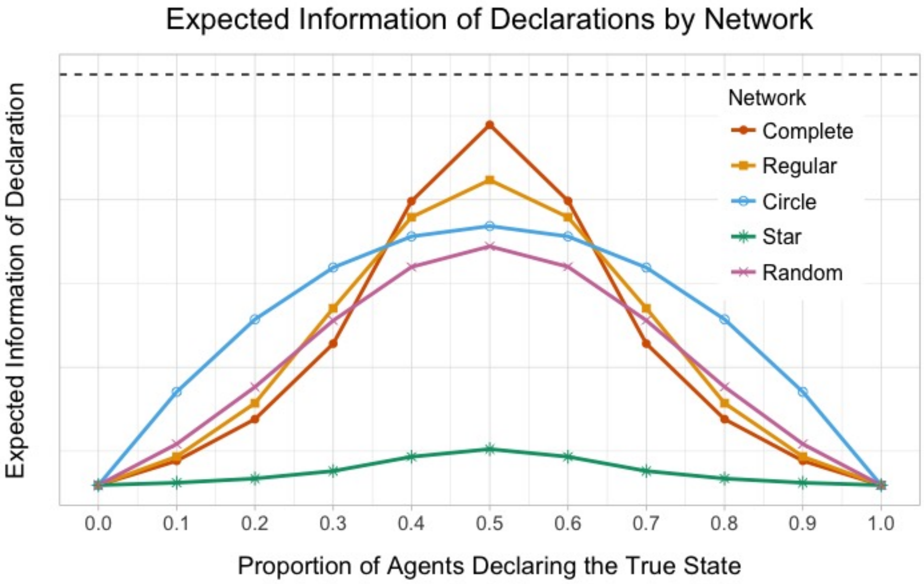


Figure 2.7: The expected informativeness of the next declaration for 10-agent networks as a function of the proportion of the population which is declaring of the true state. The dashed line denotes the expected information in the absence of any conformity.

there is an even split in declarations $N_\theta = 1/2$. Given Observation 1, it should be clear why this is so. Declarations are expected to be informative in measure to the presence of balanced dissent.

Next, we observe that, when the population is nearly split, the complete network produces the most informative declarations among the networks considered, while the circle network produces the most informative declarations when the population is near consensus. Finally, the star network provides the least informative social configuration no matter the proportion of the population making either declaration.

We may understand these results in terms of our previous insights, and sharpen them by considering large networks. On a star network, when the population is large, practically every individual has merely one neighbor. Hence, for any proportion of declarations in the population, the star network will be in the minimally informative state. That is, $\boldsymbol{I}(\mathcal{G}_{star}|N_\theta) = N_\theta \boldsymbol{I}(1) + (1 - N_\theta)\boldsymbol{I}(0) = \boldsymbol{I}(0)$.[18]

[18] Given the assumption of symmetry of expected informativeness across $N_\theta = 1/2$, we have that $\boldsymbol{I}(0) = \boldsymbol{I}(1)$, and, more generally, that $\boldsymbol{I}(1/2 - c) = \boldsymbol{I}(1/2 + c)$ for $c \in [0, 1/2]$.

Proposition 10. *For large networks, the star network is minimally informative in any state.*

On a complete network, when the population is evenly divided, $N_\theta = 1/2$, each individual is in the optimal position to make informative declarations. When all individuals are neighbors and the population is sufficiently large, the expected informativeness of the network as a whole recapitulates the expected informativeness of individual declarations given in FIGURE 2.6. That is, $\boldsymbol{I}(\mathcal{G}_{complete}|N_\theta) = \boldsymbol{I}(N_\theta)$. Given Observation 1, we show that no network can be more informative in such a state.

Proposition 11. *For large networks, when the population is evenly split in declarations, the complete network is maximally informative.*

On a circle network, when the population is near consensus, a single dissenting individual can make it possible for both her neighbors to declare their honest beliefs. That is, given that each individual has two neighbors, their neighbors' declarations are binomially distributed with the success parameter given by the population proportion of declarations, $\boldsymbol{I}(\mathcal{G}_{circle}|N_\theta) = N_\theta^2\boldsymbol{I}(0) + 2N_\theta(1 - N_\theta)\boldsymbol{I}(1/2) + (1 - N_\theta)^2\boldsymbol{I}(1)$. Contrast this with the complete network where, near consensus, every individual faces strong incentives to conform.

Proposition 12. *For large connected networks, for a range of states near consensus in declarations, the circle network is the maximally informative network.*

More generally, we can express the expected informativeness of the declaration of any individual with d connections and proportion N_θ of her neighbors declaring the true state as

$$E_{N_\theta}[\boldsymbol{I}_d] = \sum_{k=0}^{d} \binom{d}{k} N_\theta^k (1 - N_\theta)^{d-k} \boldsymbol{I}\left(\frac{k}{d}\right). \qquad (*)$$

From this, we may derive the informativeness of any network, when we conceive of networks as admixtures of proportions of individuals with different numbers of neighbors.

Given any large network, it can be represented as a distribution $\mu = \langle \mu_d \rangle$ over the degree d of individuals within the network. Thus, the expected informativeness of the network will be $\boldsymbol{I}(\mathcal{G}_\mu|N_\theta) = \sum_d \mu_d \cdot E_{N_\theta}[\boldsymbol{I}_d]$. Using this, we provide bounds for the informativeness of epistemic networks near consensus.

Proposition 13. *For large networks, near consensus, any network (including any regular or random network) of minimum degree at least two will be intermediate in informativeness between the circle and complete network.*

2.6 Conclusion

The prospects for a flourishing democracy depend crucially on our ability to engage in successful collective inquiry, discussion, and decision. We have seen that when social learning proceeds under the influence of conformity bias, the social structure of the community becomes a critical determinant of the success of public discourse. And so it behooves us to begin to identify the social structures that promote or derange the reliability of public discourse in arriving at true beliefs.

That disagreement and diversity in publicly held opinions can be optimal for honest communication gives us our key insight into understanding the effects of different social networks. This also provides a distinctive epistemic justification for the diversity of opinions: as a way to bolster informative communication when faced with the pervasive influence of conformity bias. The question as to which social networks lead to reliable beliefs then becomes a question as to which social networks produce and sustain optimal patterns of disagreement throughout the process of inquiry.

We demonstrated that, in the presence of even a modicum of conformity bias, the star network always provides the worst conditions for informative communication, the complete network provides optimal conditions exactly when the population is evenly divided, the circle network provides optimal conditions near consensus, and that, near consensus, all sufficiently connected networks will be intermediate in informativeness between the circle and complete networks.

This has implications for real-world social networks, which tend to exhibit low average degree and high degree-centrality (Watts and Strogatz, 1998). We may conjecture that, when we suspect conformity bias at play in social discourse and decision-making, interventions which reduce the density of connections of a social network while still keeping it connected, and interventions which decrease its centralization by reducing the relative influence of central individuals, may lead to more informative communication—and so to more reliable beliefs—for the epistemic community as a whole.

2.7 Mathematical Appendix

Learning from others' declarations

When agent i declares $x = \theta$, we know that it was her best response. As previously mentioned, this implies that the following condition holds:

$$\alpha_i(2P_i(\theta) - 1) + (1 - \alpha_i)(2N_i(\theta) - 1) > 0. \tag{†}$$

We plug agent i's (publicly unknown) posterior belief $P(\theta|\sigma)$ into (†) to get the elaborated condition

$$\alpha_i \left(\frac{2}{1 + \frac{1 - \bar{P}}{\bar{P}} \frac{1 - \sigma}{\sigma}} - 1 \right) + (1 - \alpha_i)(2N_i(\theta) - 1) > 0 \tag{‡}$$

where $\bar{P}$ denotes the (publicly known) prior $P(\theta|\boldsymbol{h}^t)$. We then compute the likelihood of agent i's declaration θ, given our public prior, as follows.

Let ϕ denote the left-hand term of our elaborated condition (‡), under which our agent would have declared θ, so that $\mathbb{I}[\phi > 0]$ is its indicator function. We then get the likelihood of the declaration given each possible state of the world,

$$\begin{aligned} P(x = \theta|\theta, \bar{P}) &= \int_A \int_\Sigma \mathbb{I}[\phi > 0] \mathrm{d}F_\theta(\sigma) \mathrm{d}G(\alpha), \\ P(x = \theta|\neg\theta, \bar{P}) &= \int_A \int_\Sigma \mathbb{I}[\phi > 0] \mathrm{d}F_{\neg\theta}(\sigma) \mathrm{d}G(\alpha). \end{aligned}$$

From these we obtain the posterior belief of the other agents in the network in light of agent i's declaration of θ using Bayes' rule

$$P(\theta|x = \theta, \bar{P}) = \left(1 + \frac{\int_A \int_\Sigma \mathbb{I}[\phi > 0] \mathrm{d}F_{\neg\theta}(\sigma) \mathrm{d}G(\alpha)}{\int_A \int_\Sigma \mathbb{I}[\phi > 0] \mathrm{d}F_\theta(\sigma) \mathrm{d}G(\alpha)} \frac{1 - \bar{P}}{\bar{P}} \right)^{-1}$$

which yields the new public belief.

Proof of Proposition 6. There are two states of the world θ and $\neg\theta$. Without loss of generality, suppose θ to be the true state of the world. Let $q(\boldsymbol{h}^t) = P(\theta|\boldsymbol{h}^t)$ be the public belief and $\boldsymbol{h}^t$ the history of declarations up to time t. As is well-known, the likelihood ratio

$$\ell(\boldsymbol{h}^t) \equiv \frac{1 - q(\boldsymbol{h}^t)}{q(\boldsymbol{h}^t)}$$

is a martingale conditional on θ. Let X be the finite set of declarations. For any given declaration $x \in X$,

$$\ell(\boldsymbol{h}^t, x) = \ell(\boldsymbol{h}^t)\frac{P(x|\boldsymbol{h}^t, \neg\theta)}{P(x|\boldsymbol{h}^t, \theta)}$$

and thus the martingale property follows:

$$E\big[\ell(\boldsymbol{h}^{t+1})|\theta\big] = \sum_{x \in X} \ell(\boldsymbol{h}^t, x)P(x|\boldsymbol{h}^t, \theta) = \sum_{x \in X} \ell(\boldsymbol{h}^t)P(x|\boldsymbol{h}^t, \neg\theta) = \ell(\boldsymbol{h}^t).$$

By Theorem 3(b) of Smith and Sørensen (2000), when evidence is unbounded, individuals almost surely converge in belief to the true state. □

We show that convergence in beliefs implies a convergence in declarations. In particular, we show that convergence in beliefs implies that the community's belief in the true state will be bounded from bellow over time. We then observe, using simple probabilistic arguments, that given sufficient time the community will almost surely arrive at a consensus state where all individuals are declaring the true state. Finally, we show that, having arrived at such a consensus with individual beliefs in the true state appropriately bounded from bellow, the community must remain at this consensus forever.

Proof of Corollary. 7 Let q and q' denote the public belief before and after hearing a declaration, respectively. Consider a focal agent i having received her evidence from Nature on a given turn. Let P_i denote the focal agent's posterior belief $P(\theta|\sigma, \boldsymbol{h}^t)$, and suppose that this agent declared $x = \neg\theta$. It is straightforward to show that if the population could observe the focal agent's posterior, the public belief would be precisely equal to her posterior

$$q'(\neg\theta, q, N_i(\theta), P_i) = P_i. \tag{*}$$

Let $\Pi(\cdot|\neg\theta, q, N_\theta)$ be the distribution over the focal agent's posterior belief given her declaration of $\neg\theta$, q the public belief when she selected her action, and $N_i(\theta)$ the proportion of her neighbors declaring θ. By (*) we can write

$$q'(\neg\theta, q, N_i(\theta)) = \int_0^1 P_i \, \mathrm{d}\Pi(P_i|\neg\theta, q, N_i(\theta)).$$

We can thus interpret the public belief as the public's expectations of the focal agent's posterior. As the public belief almost surely converges to certainty on the truth, for almost all trajectories of the public belief $\{q_t\}_{t=0}^{+\infty}$, for all $\epsilon > 0$ there exists a time T_ϵ such that, if $t > T_\epsilon$ then $q_t > 1 - \epsilon$. That is, there is a time after which the public belief in θ will always be at least $1-\epsilon$. Then choose $\epsilon = 1/2$.

With probability 1 at some point along the trajectory after T_ϵ all agents will be declaring θ. To see this, let λ be the probability all N agents choose declarations in sequence, each has an α sufficiently high such that they declare the state they believe to be more likely regardless of their neighbors' declarations, and they receive evidence such that their posterior assigns higher probability on θ. However small the probability λ might be, it exceeds 0. Hence, the probability that this event does not occur goes to zero as $t \to +\infty$.

Assume, for the sake of contradiction, that at some point after T_ϵ an agent goes against the consensus and declares $\neg\theta$, then her posterior must satisfy

$$P_i \leq -\frac{1-\alpha_i}{2\alpha_i} + \frac{1}{2}.$$

But then we get that $E[P_i|\neg\theta, \cdot] \leq 1/2$. That is, her belief in θ was less than $1/2$, which contradicts the fact that her belief was bounded from bellow. Hence, no agent can deviate from the consensus after time T_ϵ, and convergence in belief implies convergence in declaration. □

Lemma 14 (Monotonicity of Informativeness in Influence)**.** *The informativeness of a declaration about a state is monotonically increasing in its influence on the public belief.*

Proof. Without loss of generality, let the focal agent declare $x = \theta$. We show that the informativeness of her declaration, $H(q|q(\theta) = 1/2) - H(q|x = \theta))$, is monotonically increasing in its influence, $q(\theta|x = \theta) - q(\theta)$.

First, we unpack the definition of informativeness, temporarily omitting the assumption of the maximal entropy prior $q(\theta) = 1/2$, to get

$$\begin{aligned} H(q) - H(q(\theta|x = \theta) &= \mathrm{E}[-\ln(q(\theta|x=\theta))] - \mathrm{E}[-\ln(q(\theta))] \\ &= \mathrm{E}[\ln(q(\theta)) - \ln(q(\theta|x = \theta))] \\ &= \mathrm{E}\left[\ln\left(\frac{q(\theta)}{q(\theta|x=\theta)}\right)\right] \\ &= q(\theta)\cdot\ln\left(\frac{q(\theta)}{q(\theta|x=\theta)}\right) + q(\neg\theta)\cdot\ln\left(\frac{q(\neg\theta)}{q(\neg\theta|x=\theta)}\right) \end{aligned}$$

Now, let $A \equiv q(\theta)$ and $B \equiv q(\theta|x = \theta)$, so that $C \equiv B - A$ denotes the influence of the declaration $x = \theta$. Then we can re-write the preceding expression as

$$A \cdot \ln\left(\frac{A}{A+C}\right) + (1-A)\cdot\ln\left(\frac{1-A}{1-(A+C)}\right)$$

Taking the partial derivative with respect to influence C, and solving for when it is positive—i.e., for when informativeness is increasing—yields

$$A + C - 1 > 0 \quad \text{or} \quad B > 1/2.$$

And when $q(\theta) = 1/2$, we have that $B = q(\theta|x = \theta) \geq 1/2$, and so informativeness is monotonically increasing in influence, as desired. □

We will show that $q'(\theta, N_i(\theta)') < q'(\theta, N_i(\theta))$ whenever $N_i(\theta)' > N_i(\theta)$. From this it follows straightforwardly that, given $N_i(\theta) \in [0, 1]$, the most influential declaration occurs just when $N_i(\theta) = 0$.

To do so, consider a given focal agent i having received evidence $\sigma \sim f_\theta(\sigma)$ from Nature. Let $r = r(\sigma) \equiv P_i(\neg\theta|\sigma)$ be one minus her private belief, $G_{\neg\theta}(r)$ and $G_\theta(r)$ the conditional cdf's for r, and $g(r) \equiv \frac{dG_{\neg\theta}}{dG_\theta}(r)$ the Radon-Nikodym derivative of $G_{\neg\theta}$ with respect to G_θ.

Lemma 15. $g(r) = \frac{r}{1-r}$ *almost surely.*

Proof. If an agent updates her belief after observing r, it will remain unchanged. Thus from Bayes' theorem $r = P_i(\neg\theta|r) = \frac{g(r)}{g(r)+1}$. □

Lemma 16. *The ratio* $\frac{G_{\neg\theta}}{G_\theta}(r)$ *is strictly increasing for r in the common support of* G_θ *and* $G_{\neg\theta}$*.*

Proof. Let $r' > r$. From Lemma 15 we have that $g(r)$ is strictly increasing, hence,

$$G_{\neg\theta}(r) = \int_0^r g(x) dG_\theta(x) < g(r) G_\theta(r)$$

And thus

$$\begin{aligned} G_{\neg\theta}(r') - G_{\neg\theta}(r) &= \int_r^{r'} g(x) dG_\theta(x). \\ &> [G_\theta(r') - G_{\neg\theta}(r)] g(r) \\ &> [G_\theta(r') - G_{\neg\theta}(r)] \frac{G_{\neg\theta}(r)}{G_\theta(r)}. \end{aligned}$$

It follows that $\frac{G_{\neg\theta}(r')}{G_\theta(r')} > \frac{G_{\neg\theta}(r)}{G_\theta(r)}$. □

Proof of Proposition 8. Now, we proceed to show that $q'(\theta, N_i(\theta)') < q'(\theta, N_i(\theta))$ whenever $N_i(\theta)' > N_i(\theta)$. Define q' to be the posterior public belief, q the prior public belief, $N_i(\theta)$ the proportion of the focal agent's neighbors declaring θ, and $\Pi(\cdot|x_i, q, N_i(\theta))$ the posterior belief over the declaring agent's truth-seeking orientation $\alpha_i \in [0,1]$. Then

$$q'(\theta, N_i(\theta)) = \int_0^1 q'(\theta, N_i(\theta), \alpha_i) \mathrm{d}\Pi(\alpha_i|\theta, N_i(\theta), q).$$

For a given α_i in the support of $\Pi(\cdot|\theta, N_i(\theta), q)$, there exists a threshold $\bar{r} = \bar{r}(\alpha_i, q, N_i(\theta))$ such that the agent only selects $x_i = \theta$ if $r \leq \bar{r}$. From Bayes' theorem,

$$q'(\theta, N_i(\theta), \alpha_i) = \left(1 + \frac{1-q}{q}\frac{G_{\neg\theta}(\bar{r})}{G_\theta(\bar{r})}\right)^{-1}.$$

If $\bar{r}(\alpha_i, N_i(\theta)', q) \geq \bar{r}(\alpha_i, N_i(\theta), q)$ holds, and further holds strictly for a subset of α_i with positive posterior probability, then, by Lemma 16, $q'(\theta, N_i(\theta)') < q'(\theta, N_i(\theta))$.

It can be shown that the threshold $\bar{r}(\alpha_i, N_i(\theta), q)$ is strictly increasing in $N_i(\theta)$. This gives us that $q'(\theta, N_i(\theta)', \alpha_i) \leq q'(\theta, N_i(\theta), \alpha_i)$. Furthermore, having assumed that α_i and r take full support in $[0,1]$, we can find a neighborhood of $\alpha_i = 1$ with positive probability such that $\bar{r}(\alpha_i, N_i(\theta), q) > 0$ for all α_i in this neighborhood. Hence, in this neighborhood $q'(\theta, N_i(\theta)', \alpha_i) < q'(\theta, N_i(\theta), \alpha_i)$. □

Proof of Corollary 9. We have, from proposition 8, that $q'(\theta, N_i(\theta)') > q'(\theta, N_i(\theta))$ whenever $N_i(\theta)' < N_i(\theta)$. It follows directly that

$$\arg\max_{N_i(\theta)\in[0,1]} q'(\theta, N_i(\theta)) = 0.$$

Thus, the most influential declaration is made just when $N_i(\theta) = 0$. And we have, from Lemma 14, that this is also the most informative declaration. □

Proof of Proposition 10. On a large star network, proportion one of individuals have a single neighbor. So, for any proportion of the population declaring θ, every individual is in the minimally informative state where either $N_\theta = 0$ or 1. Hence, for all $N_\theta \in [0,1]$, and symmetric $\boldsymbol{I}$, $\boldsymbol{I}(\mathcal{G}_{star}) = \boldsymbol{I}(0) \leq \boldsymbol{I}(\mathcal{G})$ for any connected network $\mathcal{G}$. □

Proof of Proposition 11. On a complete network, every individual is neighbors with every other. Hence, the proportions of an individuals neighbors declaring θ is the same as the proportion of the population declaring θ. That is $N_i(\theta) = N_\theta$ for each i. The expected informativeness is maximized when an individual's neighbors are equally split $N_i(\theta) = 1/2$. Thus, when exactly half the population is declaring θ, the declaration of every individual in the population is at maximal expected informativeness. Hence, no other network can be more informative in this state. That is, when $N_\theta = 1/2$, $\boldsymbol{I}(\mathcal{G}_{complete}) = \boldsymbol{I}(1/2) \geq \boldsymbol{I}(\mathcal{G})$ for any connected network $\mathcal{G}$. □

To show that the circle is maximally informative near consensus, first we show that for regular networks of degree at least 2 informativeness is decreasing in degree near consensus. This implies that any regular network of degree greater than two is less informative than the circle network. We combine this with Proposition 10, which implies that networks of degree 1 are also less informative than the circle network, to show that the circle network is the maximally informative regular network. Next, using the fact that any network can be formulated as an admixture of individuals of various degrees we derive that the circle network is maximally informative near consensus.

Lemma 17. *For regular networks of degree at least 2, informativeness is decreasing in degree near consensus.*

Proof. Take the derivative of the informativeness of any regular network $\mathcal{G}_d$ of degree $d \geq 2$ with respect to the proportion of the population declaring the true state.

$$\frac{d}{dN_\theta}\left[\boldsymbol{I}(\mathcal{G}_d)\right] = \frac{d}{dN_\theta}\left[\sum_{k=0}^{d}\binom{d}{k}N_\theta^k(1-N_\theta)^{d-k}\boldsymbol{I}\left(\frac{k}{d}\right)\right].$$

Let N_θ go to 0. This makes it so only the constant terms of the derivative remain, and the expression simplifies to

$$\lim_{N_\theta \to 0^+}\frac{d}{dN_\theta}\left[\boldsymbol{I}(\mathcal{G}_d)\right] = d[\boldsymbol{I}(1/d) - \boldsymbol{I}(0)].$$

This term corresponds to the slope of the secant line connecting $\boldsymbol{I}(0)$ and $\boldsymbol{I}(1/d)$. Since $\boldsymbol{I}$ is an increasing function, this term must be decreasing in d. Thus, for networks of degree two and greater, informativeness is decreasing in degree near consensus. □

Lemma 18. *The circle is the maximally informative regular network near consensus.*

Proof. This follows from Lemma 17 and Proposition 10, which state that a regular network of degree 2 (the circle) is more informative than any network of greater degree near consensus, and that a regular network of degree 1 is less informative than any other at any state. Taken together, they imply that, near consensus, regular networks of degree two are maximally informative among regular networks. □

Proof of Proposition 12. Now, recall that any large connected network $\mathcal{G}_\mu$ can be formulated as an admixture $\mu = \langle \mu_d \rangle$ of proportions of individuals of degree $d \geq 1$, where $\sum_d \mu_d = 1$ and $\mu_d \geq 0$. The expected informativeness of any network then is a proportion-weighted sum of the expected informativeness of the individuals of each degree contained in the network. That is, $\boldsymbol{I}(\mathcal{G}_\mu | N_\theta) = \sum_d \mu_d \cdot E_{N_\theta}[\boldsymbol{I}_d]$. It follows from Lemma 18 that, near consensus, any network not entirely composed of individuals of degree two is strictly less informative than one which is in fact composed entirely of individuals of degree two. Thus, when $N_\theta = 0$ or 1, $\boldsymbol{I}(\mathcal{G}_{circle}) > \boldsymbol{I}(\mathcal{G}_\mu)$ for any $\mathcal{G}_\mu$ such that $\mu_0 = 0$ and $\mu_2 \neq 1$. □

Proof of Proposition 13. It follows directly from Lemma 17 that, near consensus, the maximally and minimally informative regular networks of degree at least two are the circle and complete network, respectively. We combine this with the fact that any large network $\mathcal{G}_\mu$ can be formulated as an admixture $\mu = \langle \mu_d \rangle$ of regular networks of degree d, and with the linearity of expected informativeness, to adduce that the informativeness of any network is bounded above by that of the circle network and bounded bellow by the complete network. That is, when $N_\theta = 0$ or 1, $\boldsymbol{I}(\mathcal{G}_{circle}) \geq \boldsymbol{I}(\mathcal{G}_\mu) \geq \boldsymbol{I}(\mathcal{G}_{complete})$ for any $\mathcal{G}_\mu$ such that $\min\{d : \mu_d > 0\} \geq 2$. □

2.8 Computational Appendix

A GUI for exploring the model is available at: `https://amohseni.shinyapps.io/Truth-and-Conformity-on-Networks/`.

The full R source code for the computational model can be found at: `https://github.com/amohseni/Truth-and-Conformity-on-Networks`

Stochastic Stability and Disagreement Between Dynamics 3

Abstract

The replicator dynamics and Moran process are the main deterministic and stochastic models of evolutionary game theory. These models are connected by a mean-field relationship—the former describes the expected behavior of the latter. However, there are conditions under which their predictions diverge. I demonstrate that the divergence between their predictions is a function of standard techniques used in their analysis, and of differences in the idealizations involved in each. My analysis reveals problems for stochastic stability analysis in a broad class of games. I also demonstrate a novel domain of agreement between the dynamics, and draw a broader methodological moral for evolutionary modeling.

3.1 Introduction

The replicator dynamics (Taylor and Jonker, 1978) and frequency-dependent Moran process (Moran, 1962) are the main deterministic and stochastic dynamics of evolutionary game theory (Cressman and Tao, 2014; García and Traulsen, 2012). Both dynamics capture the basic idea that phenotypes that are more fit than the population average tend to grow in proportion, while phenotypes less fit than average tend to shrink in proportion. The replicator dynamics gives us a deterministic description of the behavior of evolution, assumes infinite populations, and isolates the influence of selection. The Moran process gives us

	A	B
A	1	2
B	2	1

Table 3.1: A 2 × 2 symmetric anti-coordination game.

a stochastic description of evolution, assumes finite populations, and introduces the effects of drift.

Importantly, the two dynamics are connected by a mean-field relationship (Benaïm and Weibull, 2003, 2009). Intuitively, the replicator dynamics describes the expected behavior of the Moran process for large populations over finite stretches of time.[1]

Yet, there exists a striking puzzle: employing standard methods of analysis, the two dynamics can make contradictory predictions (Sandholm, 2009, Ch.12). When the interactions within a population are modeled by an anti-coordination game, or game containing an anti-coordination subgame, the replicator dynamics may predict that selection will favor polymorphism,[2] but it is said that the Moran process shows that such polymorphisms cannot persist in the long run (Taylor et al., 2004; Novak, 2007). I examine this puzzle, and show that its standard explanation is not quite right. I demonstrate that, even in the long run, there are a range of conditions under which the Moran process sustains polymorphism. Under conditions I characterize the long run behavior of the Moran process will realign with the predictions of the replicator dynamics.

The misunderstanding of the behavior of the Moran process stems from a shortcoming in a standard technique of analysis: stochastic stability analysis. And the shortcoming of stochastic stability results from its assumption of vanishing mutation rates. My results indicate problems for stochastic stability in a broad class of games, reveal a novel domain of agreement between the two dynamics, and suggest a methodological moral for evolutionary modeling.

To understand our motivating puzzle, we can consider the simple anti-coordination game given in Table 3.1, and examine the predictions as to the evolutionary outcomes of its corresponding population game under each

[1] The replicator dynamics also provides a mean field for other dynamics, such as reinforcement learning (Benaïm and Weibull, 2003), and emerges from distinct revision protocols, including pairwise proportional imitation, imitation driven by dissatisfaction, and imitation of success (Sandholm, 2009, Ch.5.4).

[2] *Polymorphisms*, here, are population states in which multiple phenotypes are present. They are contrasted with *monomorphisms*, in which only a single type is present.

dynamics. For both dynamics let us assume: large populations, random pairwise interactions, true breeding, the absence of mutation, and infinite-horizon play. Under the replicator dynamics, the prediction is that, from most all initial conditions, evolution will deliver the population to the *polymorphic state* $x = 1/2$,[3] where A-types and B-types coexist in equal proportions. In contrast, for the same anti-coordination game, the Moran process predicts that evolution will deliver the population, with equal probability, to one of the two *monomorphic states* $x = 0$ or $x = 1$,[4] where the population is composed entirely of either A-types or B-types. The evolutionary outcome that is a moral certainty in one model is an impossibility in the other.

Such a divergence in the predictions of the two dynamics leads naturally to the following questions: How are we deriving the predictions of each dynamics? And what is the cause of their divergence?[5]

The standard explanation for divergence in such cases is that the dynamics differ in the time-horizons of their predictions: the replicator dynamics approximates the short-to-medium run behavior of evolution, while the the Moran process can capture its long run behavior (Taylor et al., 2004; Novak, 2007). The prediction of the replicator dynamics is polymorphism, and this correct for the short-to-medium run. The prediction of the Moran process is monorphism, and this is correct for the long run. Young (1998, 47) states this clearly: "While [the replicator dynamics] may be a reasonable approximation of the short run (or even medium run) behavior of the process, however, it may be a very poor indicator of the long run behavior of the process."

The dynamics differ with respect to the time-horizons of their predictions. This is true, but in the case of interest, this is not the cause of the divergence in their predictions, and it is not the answer to our puzzle. The cause of the divergence lies in the standard technique employed to derive predictions from the Moran process, *stochastic stability analysis*, introduced to game theory by Foster & Young (1993). Under conditions I will characterize, stochastic stability leads to the mis-prediction of homogeneity where long run diversity is to be expected.

Why does this matter? In brief, because the technique of stochastic stability analysis is ubiquitous. Among those having deployed stochastic stability in

[3] For the replicator dynamics, these will be *asymptotically stable* states. These will be explained in §2.1.

[4] For the Moran process, these will be either *absorbing* states or *stochastically stable* states, depending on the presence of mutation. These will be explained in §2.2 and §2.3.

[5] Another important—and open—question is: how, generally, should we meaningfully compare the predictions of stochastic and deterministic dynamics?

	A	B
A	a	b
B	c	d

Table 3.2: A 2×2 symmetric game

the analysis of the Moran process, and related processes,[6] include: Binmore & Samuelson (1995; 1997), Fudenberg & Imhof (2004; 2006), Fudenberg et al (2006), Imhof et al (2006), Nowak et al (2004; 2007), Ohtsuki et al (2007), Sandholm (2007; 2009; 2010; 2012), Taylor et al (2004), Trauelsen & Hauert (2010), and Young (1993; 1998; 2005; 2015). Evolutionary game theorists use stochastic stability analysis to explore and explain various phenomena in the domains of cultural and biological evolution, ranging from the diffusion of innovations to the emergence of conventions. Stochastic stability is a standard tool in both theoretical and applied work. Given this, understanding its limitations is important.

The structure of this paper is as follows. In §2, I will introduce the replicator dynamics and Moran process models along with the concepts of asymptotic stability, replacement probabilities, and stochastic stability needed to understand our results. In §3, I will demonstrate the conditions under which stochastic stability will mis-predict the long run behavior of the Moran process, polymorphisms will persist, and the behavior of the replicator dynamics and Moran process will realign. In §4, I will discuss real-world applications where one can expect my results will matter, and suggest a methodological moral for evolutionary modeling. In §6, I conclude.

3.2 The Dynamics

The Replicator Dynamics

The replicator dynamics is the "first and most important model of evolutionary game theory" (Cressman and Tao, 2014, 1081). This is due to the fact that it allows us to isolate the qualitative influence of selection on evolution, unperturbed by the complicating factors of mutation, drift, recombination, and so on. The leading idea behind the replicator dynamics is that types that are more fit than the population average fitness grow in relative proportion, and types

[6] These include the closely related Markov processes of Fermi and Wright-Fisher.

that are less fit than average shrink in proportion. This can be described by a system of differential equations[7]

$$\dot{x}_i = x_i[u(i, x) - u(x, x)] \quad \text{for } i \in S$$

where S is the set of possible types, $\dot{x}_i$ denotes the rate of change of the population proportion of type i, x_i denotes the population proportion of type i, $u(i, x)$ denotes the expected fitness for type i from interacting with the population, and $u(x, x)$ denotes the population average fitness.

We derive predictions from the replicator dynamics by finding the *asymptotically stable states* of the dynamics for a given game, and equating these with the plausible outcomes of evolution for that game.[8] A population state is asymptotically stable just in case it is both *stable* and *attracting*. Intuitively, a state is stable if states near it remain near it, and attracting if states near it tend toward it. This gives us our prediction of the behavior of a process described by the replicator dynamics.

For the simple class of 2×2 symmetric games under the replicator dynamics, five qualitatively distinct outcomes are possible. These can bee seen in Table 3.3. Our puzzle concerns the class of anti-coordination games, shown in the third row of the table, and labeled the 'polymorphic case'. This is where we find polymorphisms that are asymptotically stable under the replicator dynamics. Anti-coordination games constitute and important class of interaction structures, and have been used in explanations of ritualized animal conflict (Maynard Smith, 1974), sex ratios (Hamilton, 1967), and bargaining norms (Skyrms, 1996).

We derive the prediction of the replicator dynamics for anti-coordination games (Table 3.3, third row) as follows. We solve for the fixed points of the dynamics, where the rate of change in population proportions of each type is zero, i.e., $\dot{x}_1 = \dot{x}_2 = 0$. This yields three states: the two monomorphic states composed entirely of one type or the other, and a polymorphic state, i.e.,

[7] The Replicator dynamics can also be formulated for discrete time—the Maynard-Smith formulation (1982)—by a system of difference equations, which, under some conditions, yield subtly different results from their continuous time counterpart (Cressman, 2003). However, for 2×2 games, the qualitative predictions of the two formulations coincide. Thus, here, without loss of generality, I will work with the continuous time formulation exclusively.

[8] Asymptotic stability does not exhaust the plausible outcomes of the replicator dynamics. In more complex games, disequilibrium behavior such as cycles and strange attractors, along with sets of collectively but not individually stable states, will not be asymptotically stable but may still constitute plausible outcomes of the dynamics. For a survey and analysis of this issue, see (Mohseni, 2017). However, for the class of 2×2 symmetric games considered here, there is a one-to-one correspondence between evolutionarily significant outcomes and asymptotically stable states. So, we can proceed comfortably with asymptotic stability as our stability concept for the replicator dynamics.

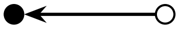
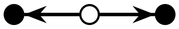
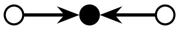
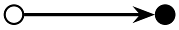
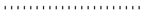
Game Type	Payoffs	Phase Portrait	Asymptotically Stable States
A dominates B	$a > c$ $b > d$	A  B	All-A state
Bi-stable case	$a > c$ $b < d$	A  B	All-A and All-B states, with basins of attraction divided at $\frac{d-b}{d+a-c-b}$ A-types
Polymorphic case	$a < c$ $b > d$	A  B	A mixed state, with $\frac{b-d}{b+c-a-d}$ A-types
B dominates A	$a < c$ $b < d$	A  B	All-B state
Neutral case	$a = c$ $b = d$	A  B	None

Table 3.3: 2×2 symmetric games under the replicator dynamics. Opaque circles denote asymptotically stable states, empty circles denote unstable fixed points, dotted lines denote sets of unstable fixed points, and arrows indicate the direction of selection.

$\{0, \frac{b-d}{b+c-a-d}, 1\}$. We assess the stability of these states by examination of the eigenvalues of Jacobian matrix for the dynamics, which reveals that only the mixed state is asymptotically stable. Given this, we know that a population starting at the polymorphism with proportion $\frac{b-d}{b+c-a-d}$ A-types will remain there, and that, from most all initial conditions,[9] the dynamics will converge to the polymorphism.

The Frequency-Dependent Moran Process

The Moran process is a birth-death process in which, for each time step, two individuals are chosen: one for reproduction and the other for elimination. The individual chosen for birth is determined, probabilistically, by the relative fitness of the types within the population, and the individual chosen for death is selected at random. So, if we consider a population of N individuals whose payoff from interaction are described by Table 3.2, then the fitnesses f_i, g_i of the types A, B can be described as functions of the number i of A-types,

$$f_i = 1 - w + w\frac{a(i-1) + b(N-i)}{N-1} \quad \text{and} \quad g_i = 1 - w + w\frac{ci + d(N-i-1)}{N-1},$$

where w denotes the intensity of selection, or the game's contribution to the net fitness of the organism. Observe that $w = 1$ implies that an individual's fitness

[9] Initial conditions in which some proportion of each type is present in the population.

is entirely determined by her interactions in this game, and $w = 0$ implies that the game makes no contribution to her fitness.

Individuals reproduce at a rate proportional to their fitness. The rate of reproduction then for A -types is if_i and for B-types is $(N-i)g_i$. Each period, one offspring is chosen at random to enter the population. So, the probability of adding an A-type offspring is $\frac{if_i}{if_i+(N-i)g_i}$, and the probability of adding a B-type offspring is $\frac{(N-i)g_i}{if_i+(N-i)g_i}$. After reproduction, one individual is chosen at random to be removed from the population, so that with probability $\frac{i}{N}$ an A-type is removed, and with probability $\frac{N-i}{N}$ a B-type is removed. This makes it so that the population size remains constant.

Formally, we define the Moran process with population size N as a Markov process $\{X_t^N\}$ over the finite state space $\chi = \{1, \ldots, N\}$ of possible population states, with transition probabilities between states given by

$$P_{i,j} = \begin{cases} \dfrac{N-i}{N}\dfrac{if_i}{if_i+(N-i)g_i}, & \text{if } j = i+1 \\ \dfrac{i}{N}\dfrac{(N-i)g_i}{if_i+(N-i)g_i}, & \text{if } j = i-1 \\ 1 - P_{i,i+1} - P_{i,i-1}, & \text{if } j = i \\ 0, & \text{otherwise,} \end{cases}$$

which composes a tri-diagonal matrix. Note that $P_{0,0} = P_{N,N} = 1$, so that the process has two absorbing states, $i = 0$ and $i = N$, and that all other states are transient. An absorbing state is a state that, once visited by the process, is never escaped. A transient state then is one which will only be visited a finite number of times before the process arrives at some absorbing state. Note that, in the limit of time, with probability one, the process will reach one or the other absorbing state.

For the simple class of 2×2 symmetric games under the Moran process, as with the replicator dynamics, we can examine five distinct cases (Table 3.4) when the population size is large.[10] For each case, the outcomes are described in terms of the relative probability of arrival of the process at each of the absorbing states. In particular, we compare the probability of a single mutant coming to replace the incumbent type, and take over the population. This yields the *replacement probabilities*[11] ρ_{AB} and ρ_{BA}, where ρ_{AB} denotes the probability

[10] We take the large population limit for the Moran process to allow for meaningful comparison with the replicator dynamics. For analysis of the changes in the behavior of the Moran process as a function of population size see (Taylor et al., 2004).

[11] For an exposition of the details of this approach, see (Nowak et al., 2004).

Game Type	Payoffs	Replacement Probabilities	Description
A dominates B	$a > c$ $b > d$	$\rho_{BA} < \frac{1}{N} < \rho_{AB}$	Selection opposes B and favors A.
Bi-stable case	$a > c$ $b < d$	Not $(\frac{1}{N} < \rho_{BA}, \rho_{AB})$	Selection may favor A or B, but not both.
Polymorphic case	$a < c$ $b > d$	Not $(\rho_{BA}, \rho_{AB} < \frac{1}{N})$	Selection may oppose A or B, but not both.
B dominates A	$a < c$ $b < d$	$\rho_{AB} < \frac{1}{N} < \rho_{BA}$	Selection opposes A and favors B.
Neutral case	$a = c$ $b = d$	Not $(\frac{1}{N} < \rho_{BA}, \rho_{AB})$, or $\rho_{BA} = \rho_{AB} = \frac{1}{N}$	Selection favors A or B depending on the sign of $(a+b) - (c+d)$, or is neutral if $a + c = b + d$.

Table 3.4: 2×2 symmetric games under the Moran process with large populations.

of a single A-type individual leading to the takeover of an otherwise B-type population, and ρ_{BA} denotes the probabilities of the inverse process. The replacement probabilities of types are compared to those of a neutral mutant (where $a = b = c = d$), which will come to fixation with probability $1/N$. We say that *selection favors* a type if its replacement probability is greater than that of a neutral mutant, and that *selection opposes* a type if its replacements probability is less than that of a neutral mutant.

We derive the predictions of the Moran process for anti-coordination games (Table 3.4, third row) by calculating th replacement probabilities for each type. This yields three possibilities: $\rho_{BA} < \frac{1}{N} < \rho_{AB}$, $\rho_{AB} < \frac{1}{N} < \rho_{BA}$, or $\frac{1}{N} < \rho_{AB}, \rho_{AB}$. That is, either selection favors one type replacing the other, or it favors both replacing one another.[12] What we see is that, for anti-coordination games, selection must favor at least one type in coming to dominate the population. In the absence of mutation, polymorphism is temporary, and evolution inevitably attains homogeneity.

The Frequency-Dependent Moran Process with Mutation

With the introduction of mutation the behavior of the Moran process changes qualitatively. Absorbing states disappear, and there is positive probability that the process will transit within finite time from any given state to any other. Thus, in the limit of time, the process visits each state infinitely often. Since absorption will not occur, replacement probabilities are no longer appropriate,

[12] In such cases, one can examine the sign of $\rho_{AB} - \rho_{BA}$ to determine which type is more or less favored by selection.

and a different method of analyzing the behavior of the process is needed. This method is to find the long run distribution of time spent by the process over the possible population states.

Formally, we define the Moran process with population size N and mutation rate η as an ergodic process[13] $\{X_t^{N,\eta}\}$ over the finite state space $\chi = \{1, \dots, N\}$, with transition probabilities given by

$$\hat{P}_{i,j} = \begin{cases} (1-\eta)\dfrac{N-i}{N}\dfrac{if_i}{if_i+(N-i)g_i} + \eta\dfrac{N-i}{N}\dfrac{(N-i)g_i}{if_i+(N-i)g_i}, & \text{if } j = i+1 \\ (1-\eta)\dfrac{i}{N}\dfrac{(N-i)g_i}{if_i+(N-i)g_i} + \eta\dfrac{i}{N}\dfrac{if_i}{if_i+(N-i)g_i}, & \text{if } j = i-1 \\ 1-\hat{P}_{i,i+1}-\hat{P}_{i,i-1}, & \text{if } j = i \\ 0, & \text{otherwise} \end{cases}$$

where $\hat{P}_{0,0} = \hat{P}_{N,N} = 1 - \eta$. Note the mutation terms. What they capture is that, most of the time $(1-\eta)$, selection behaves as normal, but in a minority of instances η, an offspring that was to be an A-type will become a B-type, and vice versa.[14]

Now, to understand the long term behavior of the Moran process we can compute its stationary distribution which captures proportion of time spent at each population state. Formally, a probability distribution $\mu \in \mathbb{R}^\chi$ is a *stationary distribution* of the ergodic process $\{X_t^{N,\eta}\}$ if

$$\sum_{i\in\chi} \mu_i P_{i,j} = \mu, \qquad \text{for all } j \in \chi.$$

That is, a stationary distribution is a probability vector such that taking its product with the matrix of transition probabilities simply returns itself.

We know that such a distribution exists, since every ergodic process has a unique stationary distribution, and that it is history independent. That is, from any initial distribution over states, the distribution of time spent by the process over population states converges to that given by stationary distribution.

Typically, however, we do not derive predictions from the Moran process by finding its stationary distribution. This is due to the fact that general analytic

[13] An ergodic process is a Markov process that is both irreducible (every state is reachable from any other), and aperiodic (the greatest common divisor for the number of steps to return to each state is one). It is easily verified that the Moran process with mutation is indeed ergodic.

[14] Here, we have assumed that mutation is symmetric, but it need not be so. Asymmetric mutation can be accounted for by formulating the rate of mutation for one type as a ratio of the other $r\eta$, where r is a positive constant. For an analysis of the affects of asymmetric mutation rates see (Traulsen and Hauert, 2010).

Game Type	Payoffs	Stochastically Stable States	Description
A dominates B	$a > c$ $b > d$	$\mu_A \to 1$	The stationary distribution is a point-mass at the all-A state.
Bi-stable case	$a > c$ $b < d$	$\mu_A \to 1$ xor $\mu_B \to 1$	The stationary distribution is a point-mass at either the all-A or all-B state.
Polymorphic case	$a < c$ $b > d$	$\mu_A \to 1$ xor $\mu_B \to 1$	The stationary distribution is a point-mass at either the all-A or all-B state.
B dominates A	$a < c$ $b < d$	$\mu_B \to 1$	The stationary distribution is a point-mass at the all-B state.
Neutral case	$a = c$ $b = d$	$\mu_A \to 1$ xor $\mu_B \to 1$ xor $\mu_A, \mu_B \to \frac{1}{2}$	The stationary distribution is a point-mass at all-A or all-B depending on the sign of $(a + b) - (c + d)$, or evenly split between the states if $a + c = b + d$.

Table 3.5: 2 × 2 symmetric games under the Moran process with mutation and large populations.

forms of the stationary distribution for the Moran process for complex games are not known,[15] and because the stationary distribution applies positive probability to every state, as opposed to yielding unique predictions (Harper and Fryer, 2016).

Instead, the drive for analytic tractability and unique equilibrium prediction motivates the use an alternative: stochastically stability analysis. Stochastically stable states are just those that retain mass in the stationary distribution when we take the limit as mutation approaches zero.[16] Formally, a state $i \in \chi$ is *stochastically stable* if

$$\lim_{\eta \to 0} \mu_i^{N,\eta} > 0.$$

We saw that, for the Moran process, in the absence of mutation, all and only monomorphic states were absorbing states. Now, with vanishing mutation, the stationary distribution collapses (typically) to a point-mass on just one of these absorbing states. As mutation vanishes, the behavior of the ergodic process approaches that of the absorbing chain, and so spends most of its time near one or another monomorphic state.

[15] The exceptions to this are for 2 × 2 games under arbitrary revision protocols, and for potential games under exponential revision protocols (Sandholm, 2009).

[16] Stochastic stability is often solved for using particular well-chosen graphs that capture the difficulty of transitioning from each absorbing state (of the original absorbing chain) to any other. For a presentation of the relevant techniques, see Ch. 3.2 of Young (1998). I present a more general formulation better suited to my project.

We derive the predictions of the Moran process with mutation for anti-coordination games (Table 3.5, third row) by finding which states retain positive mass in stationary distribution as mutation vanishes.[17] This yields two possibilities: $\mu_A \to 1$, or $\mu_B \to 1$.[18] That is, either the all-A or all-B state, but not both, can be stochastically stable. Once again, polymorphism cannot be selected.

3.3 The Long Run Persistence of Diversity

Why can't stochastically stable states be polymorphisms? This is by construction: a stochastically stable state of an ergodic process will be an absorbing state of its corresponding absorbing chain. Stochastic stability is defined for an ergodic process, and is determined by identifying the states that retain mass in the stationary distribution of the process as mutation vanishes. As mutation vanishes, the behavior of the ergodic process approaches that of the absorbing chain. Polymorphisms cannot be absorbing states, and thus cannot be stochastically stable.

In most game types, the qualitative predictions of asymptotic stability for the replicator dynamics and the predictions of stochastic stability for the Moran process are in basic agreement (see Tables 3.3, 3.4, 3.5). In the case of coordination games and dominating strategy games (rows 1, 2, 4), for large populations, the asymptotically stable state with the largest basin of attraction has the greatest replacement probability and is uniquely stochastically stable. Predictions differ in the in the polymorphic case (compare row 3 of Tables 3.3, 3.4, 3.5), and the neutral case (compare row 5 of Tables 3.3, 3.4, 3.5). The latter is to be expected, as the replicator dynamics explicitly abstracts away from the effects of drift. Disagreement about the polymorphic case is more puzzling. For an anti-coordination game, polymorphism is uniquely asymptotically stable under the replicator dynamics, but only monomorphic states can be stochastically stable under the Moran process.

The standard explanation we have seen accounts for this divergence in predictions by positing that the replicator dynamics fails to capture the long run behavior of the Moran process. The Moran process will, due to stochasticity, eventually arrive at an absorbing state of the process, where it will spend most of its time, trapped by low mutation rates. The following excerpt from Sandholm (2009, 208) tells this story:

[17] See (Fudenberg and Imhof, 2004) for the relevant technique.

[18] In the knife-edge case where $a = d$ and $b = c$ we get $\mu_A, \mu_b \to 1/2$.

	A	B
A	1	3
B	2	1

Table 3.6: A 2 × 2 anti-coordination game.

> "The stochastic process typically moves in the direction indicated by the mean dynamic. If the process begins in the basin of attraction of a rest point or other attractor of this dynamic, then the initial period of evolution generally results in convergence to and lingering near this locally stable set. ... However, [since the process is irreducible] this cannot be the end of the story. Indeed, the process eventually reaches all states, and in fact visits all states infinitely often. This means that the process must leave the basin of the stable set visited first; it then enters the basin of a new stable set, at which point it is extremely likely to head directly to the set itself. The evolution of the process continues in this fashion, with long periods near each attractor punctuated by sudden jumps between them."

Young (1998, 20) calls this the "punctuated equilibrium effect," echoing that, "When the stochastic shocks are small, the mode of this frequency distribution [the stationary distribution] will tend to be close to the stochastically stable [states] predicted by the theory."

But this characterization turns out to be insufficient for the polymorphic case. How small must the mutation rate be? And what role do population size and intensity of selection play? In anti-coordination games under the Moran process, the population may indeed spend the majority of its time at or near a polymorphic equilbrium, even in the long run. This occurs when there is small but non-vanishing mutation, and sufficiently large population size and intensity of selection.

To illustrate the potential divergence of the actual and predicted behaviors of the Moran process, consider the anti-coordination game given by Table 3.6. We fix the following parameter settings: population size $N = 100$, mutation rate $\eta = 0.01$, and intensity of selection $w = 0.2$. Now, consider the predictions of each of our stability concepts: The replacement probabilities are $\rho_{AB} \approx 0.1644$ and $\rho_{BA} \approx 0$, so selection favors A and opposes B; absorption into the all-A state is the most probable outcome of the process. Stochastic stability analysis yields that $\mu_A = 1$, so all-A is the unique stochastically stable state; in the long run, the process will spend almost all of its time near the all-A state.

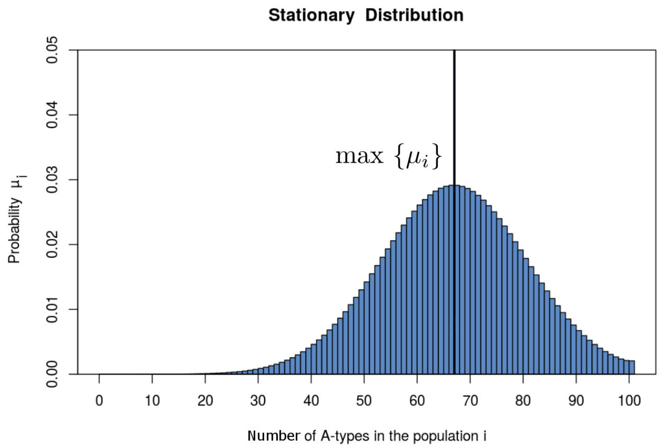


Figure 3.1: The stationary distribution for the Moran process with: population size $N = 100$, mutation rate $\eta = 0.01$, and intensity of selection $w = 0.2$. The vertical line marks the mode of the distribution.

Finally, asymptotic stability analysis yields the unique asymptotically stable state $x^* = 2/3$; from all mixed initial conditions the population will converge to a polymorphism where 2/3 of the population are A-types and 1/3 are B-types.

To see whether these predictions hold up, I analytically derive the actual long run behavior of the Moran process by calculating its stationary distribution,[19] without vanishing mutation, but rather with fixed parameter values of mutation rate.

This stationary distribution is plotted in Figure 3.1. Notice that the mode of the stationary distribution is at the state where there are 67 A-types in the 100-individual population. The process spends the most time precisely at the polymorphic state predicted by asymptotic stability under the replicator dynamics, and not at the all-A state that is stochastically stable.

To get a feel for the medium run behavior of the Moran process, we can simulate several dozen individual population trajectories, starting from random initial conditions, and evolving over a thousand birth-death events. This is

[19] In the simple case of 2×2 games, we can obtain an explicit formula for the stationary distribution: $\mu_k = \mu_0 \prod_{i=1}^{k} \frac{\hat{P}_{i-1,i}}{\hat{P}_{i,i-1}}$ for $k \in \{1, \ldots, N\}$, and $\mu_0 = \left(\sum_{k=1}^{N} \prod_{i=1}^{k} \frac{\hat{P}_{i-1,i}}{\hat{P}_{i,i-1}} \right)$, where the empty product equals one. This can also be verified, computationally, using the Chapman-Kolmogorov equation, $P^t = (P)^t$, which says that the nth-step transition matrix for a Markov process is equal to the first-step transition matrix raised to the nth power. For very large t, this can be used to approximate the stationary distribution of a given Markov process (Karlin and Taylor, 2012).

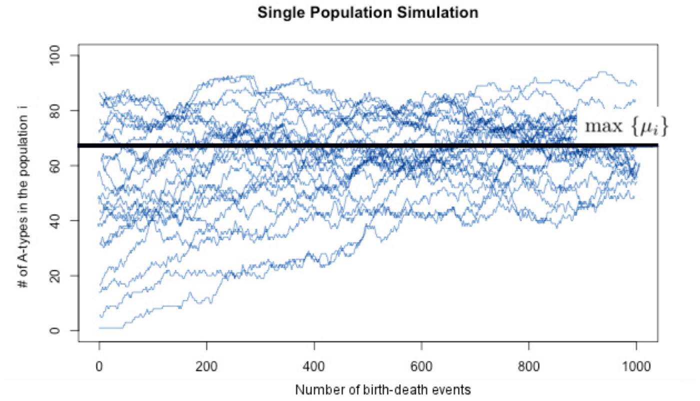


Figure 3.2: Plot of 35 population trajectories for the Moran process over 1000 birth-death events. The horizontal bar corresponds to the peak of the stationary distribution.

plotted in Figure 3.2. Again, it is clear that asymptotic stability under the replicator dynamics gives us a more accurate prediction of the behavior of the Moran process.

What we see is stochastic stability mis-predicting the actual behavior of the Moran process under particular conditions. But we want a more general characterization of when this will occur. To obtain this, we turn first to the case where there is no selection $w = 0$. Here, I use the detailed balance conditions of ergodic processes (Karlin and Taylor, 2012) to deduce when the mass of the stationary distribution will be increasing toward the center of the state space. That is, the conditions under which the peak of the stationary distribution will be at a polymorphic state. This is captured by the following lemma.

Lemma 1. *For any 2×2 game under the Moran process, in the absence of selection $w = 0$, the strong mutation condition $\eta(N + 2) > 1$ is necessary and sufficient for the mode of the stationary distribution to be a polymorphic state, and any polymorphic mode will be at the midpoint of the state space.*

What I am calling the 'strong mutation condition' corresponds to when the expected number of mutants entering a population in $N + 2$ birth-death events is greater than 1. Intuitively, here strong mutation gives us when either the population is sufficiently large such that the process rarely arrives at monomorphic states, or the mutation rate is sufficiently high such that, when the population does arrive at monomorphic states, it does not spend too much time there.

(a) $\eta(N+2) > 1$ (b) $\eta(N+2) = 1$ (c) $\eta(N+2) < 1$

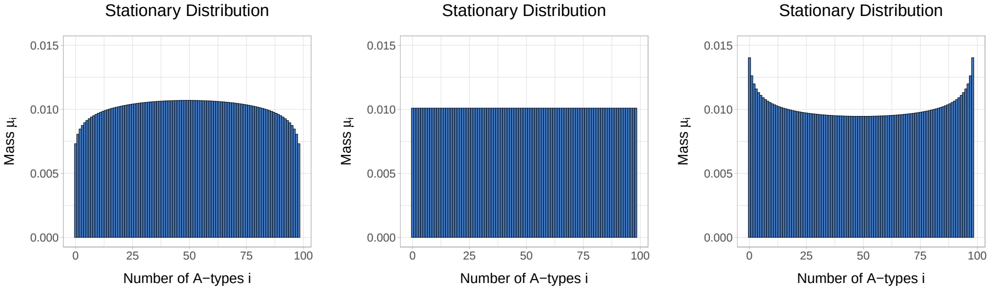


FIGURE 3.3: In the absence of selection pressures $w = 0$, satisfaction of the *strong mutation condition* $\eta(N+1) > 1$ determines the shape of the stationary distribution.

Note that strong mutation is both necessary and sufficient for the stationary distribution to exhibit a polymorphic peak (Figure 3.3). That is, just when $\eta(N+2) > 1$, the stationary distribution is concave, and climbs gradually toward its peak near the middle point $\frac{N+1}{2}$ from either side of the state space (Figure 3.3a). When $\eta(N+2) = 1$, the stationary distribution is uniform (Figure 3.3b). When $\eta(N+2) < 1$, the stationary distribution is convex, and climbs outward from its nadir near the middle point toward its peaks at the monomorphic states 0 and N (Figure 3.3c).

We can use this insight as we turn to consider the case of nonzero intensity of selection $w > 0$. Consider the dynamics of an anti-coordination game characterized by the payoffs $a < c$ and $b > d$. It may be intuitive that, when the stationary distribution is already increasing in mass toward a polymorphic state, the addition of selection pressure toward an interior equilibrium will continue to produce an interior mode. This is the essential insight from which I will derive my main result.

Before doing so, however, there are two stipulations that need to be made. First, I follow Taylor et al. (2004) in requiring that a coordination game, characterized by $a < c$ and $b > d$, further satisfy the condition that $b - d > \frac{a-d}{N} > a - c$ for finite populations. This is because, for finite populations, the qualitative dynamics of a game can be affected by the anti-correlation produced by individuals not interacting with themselves. Anti-correlation can alter the qualitative dynamics of the game. Indeed, in sufficiently small populations, each of our four game types can, in principle, be transformed into a different game.

To see why this is so, consider the case with a population composed of two individuals $N = 2$. Here, the process has three possible population states: two A-types, two B-types, and one of each type. Since transition out of each monomorphic state occurs solely via mutation, only the state where there is one of each type involves selection. In this state, only the difference of the values in the off-diagonal of the payoff matrix, $b - c$, matters. If $b - c > 0$, then A dominates B. If $b - c < 0$, then B dominates A. The game is no longer, qualitatively, an anti-coordination game.

To correct for this, we require that payoffs further satisfy $b - d > \frac{a-d}{N} > a - c$, ensuring that the game retains the qualitative dynamics of anti-coordination.[20] When N grows large, this condition is easily satisfied, and the qualitative dynamics are once again determined by the signs of the differences of the values of the column vectors, $a - c$ and $b - d$, just as with the replicator dynamics.

Second, I must also stipulate that mutation rates be reasonable: $\eta < 1/2$. It should be clear why this is so. If it is more probable that birth events are produced by mutation than by selection, then the fitnesses of the types will be reversed, and we will once again be playing a different game; a coordination game, in fact.

With these two stipulations in hand, we can turn to a sufficient condition for the persistence of diversity of types under the Moran process.

Theorem 1. *For any 2×2 symmetric anti-coordination game under the Moran process $a < c, b > d$, and $b - d > \frac{a-d}{N} > a - c$, for any intensity of selection $w > 0$ and mutation $\eta < 1/2$, when the strong mutation condition $\eta(N + 2) > 1$ is satisfied, the mode of the stationary distribution will be at a polymorphic state located between the critical point $i^* = \frac{N(b-d)+d-a}{b+c-a-d}$ and the midpoint of the state space $\frac{N+1}{2}$.*

What we have is that, when the strong mutation condition is satisfied, the peak of the stationary distribution is guaranteed to be between the midpoint of the state space and a critical point $i^* = \frac{N(b-d)+d-a}{b+c-a-d}$ that rapidly approaches the asymptotically stable state of the same game under the replicator dynamics $x^* = \frac{b-d}{b+c-a-d}$ as N grows large.[21]

What remains is for us to confirm that, as intensity of selection increases, the peak of the stationary distribution will move toward the critical point. We can answer this in the affirmative.

[20] See (Taylor et al., 2004) for a characterization of qualitative dynamics of finite games.

[21] This is clear when we state the critical point in terms of a population proportion $\frac{i^*}{N} = \frac{b-d+\frac{(d-a)}{N}}{b+c-a-d}$.

Corollary 1. *For any* 2×2 *symmetric anti-coordination game under the Moran process* $a < c, b > d$, *and* $b - d > \frac{a-d}{N} > a - c$, *for any mutation* $\eta < 1/2$, *when the strong mutation condition* $\eta(N+2) > 1$ *is satisfied, for any intensities of selection* w, w' *such that* $0 < w < w' < 1$, *the stationary distribution under* w' *puts greater mass on the states nearest the critical point* $i^* = \frac{N(b-d)+d-a}{b+c-a-d}$ *than does the stationary distribution under* w.

This is good. I note, however, that strong mutation provides *sufficient*, and *not* necessary, conditions for a polymorphic mode of the stationary distribution. An anti-coordination games can fail to satisfy the strong mutation condition, but have it so that a peak of its stationary distribution is at a polymorphic state. Strong mutation ensures that the stationary distribution increases monotonically toward some interior equilibrium, and thus ensures that there are no other peaks at the monomorphic states. For anti-coordination games where $\eta(N+2)$ is slightly less than one, but where selection pressure is great $a \ll c$ or $b \gg d$, the highest peak of the stationary distribution may still be a polymorphic state, with other smaller peaks at the monomorphic states.

In sum, when strong mutation obtains for an anti-coordination game, we know—with certainty—that stochastic stability analysis will mis-predict a monomorphic outcome when polymorphism is to be expected. But, when strong mutation does not obtain, there is still the possibility of mis-prediction.

To get an idea of the conditions under which games will exhibit polymorphic modes near the replicator dynamics prediction, we can examine the peak of the stationary distribution of a representative anti-coordination game (Table 3.6) for different values of N, η, and w.

In Figure 3.4, the darkness of each point in a plot encodes the distance, in terms of population proportions, between the peak of the stationary distribution, and the replicator dynamics prediction. In the plots, population sizes $N \in \{2, 3, \ldots, 100\}$ vary along the x-axes, mutation rates $\eta \in \{0, 0.01, \ldots, 0.5\}$ vary along the y-axes, and intensities of selection $w \in \{10^{-2}, 10^{-1}, 1\}$ vary between plots.

This accords with what we have learned so far, and illustrates our results. Where strong mutation obtains (in the space above the black curves), the peak of the stationary distribution is near the replicator dynamics prediction x^*. When the intensity of selection is low $w = 10^{-2}$, the demarcation is quite precise. As intensity of selection increases $w = 10^{-1}$, a growing range of population sizes and mutation rates (just beneath the black curves) will be compatible with an interior mode near x^*. When intensity of selection is at its maximum $w = 1$, strong mutation will continue to provide a sufficient, but not necessary, condition for polymorphism.

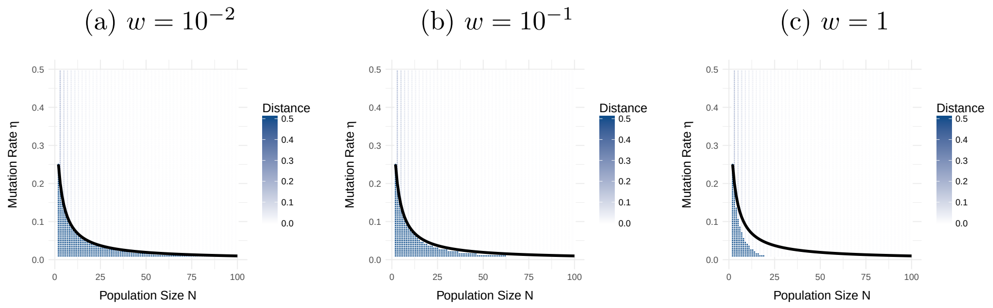

(a) $w = 10^{-2}$ (b) $w = 10^{-1}$ (c) $w = 1$

Figure 3.4: Distances of the mode of the stationary distribution from the replicator dynamics prediction for different values of N, η, w. The curve in black corresponds to $\eta(N+1) = 1$, above which the strong mutation condition is satisfied.

3.4 Discussion

I have characterized conditions under which we can anticipate the behavior of the Moran process will be mis-characterized by stochastic stability, and where long term diversity will persist. But should we expect these conditions to obtain in nature? If, indeed, the strong mutation condition were never to be satisfied, then we might comfortably rely on stochastic stability analysis without fear of it leading us astray. To see if this is so, we can survey representative population sizes, mutation rates, and intensities of selection from relevant real-world evolving populations.

Considering the canonical case of *E. coli* bacteria, we have that the per-site mutation rates are typically of the order of 10^{-4} mutations per allele per replication (Tenaillon et al., 2016). That is, an expected 1 out of $5,000$ bacteria produced carry at least one mutation at a locus of interest. Typical bacterial populations, however, are of the order of $10^6 - 10^8$. This yields a mutation strength of $\eta(N+2) \approx 20,000$. That is, there will be an average of twenty thousand mutations per population per generation. This is deep into the territory of strong mutation. Moreover, the population sizes and mutations rates of many bacteria are comparable (Drake et al., 1998). Bacterial populations, it seems, will exhibit population sizes and mutation rates that suggest their long run evolutionary behavior will typically be at odds with the predictions of stochastic stability.

Humans, on other hand, exhibit per-allele mutation rates ranging from 10^{-5} to 10^{-10} (Drake et al., 1998). Throughout much of our evolutionary history, *H. sapiens* subsisted in hunter-gatherer groups averaging 50 to 150 individuals (Bowles and Gintis, 2011). Hence, human biological evolution will typically not satisfy strong mutation. The same will be true of many complex organisms, such as mammals (Kumar and Subramanian, 2002).

However, in the case of cultural evolution, we expect mutation rates—or noise in the transmission of behavior via social learning and imitation—to be potentially much higher (Boyd and Richerson, 1985). Taking the example of humans, for a group of 100 individuals, an innovation or error rate in behavior transmission of just over 10% would satisfy strong mutation. Given that the Moran process is often used to model processes of cultural evolution, it will be important to know when an evolutionary process satisfies the strong mutation condition.

We should note that the relationship between strong mutation and persistent polymorphism is not guaranteed to hold in other game structures. We can expect the analysis will vary for extensive-form games, with more players and strategies, and with the introduction of social or spatial structure, an so on. But it is reasonable to imagine that qualitatively similar conditions may hold for other classes of games. The question as to the limits of the agreement of the dynamics for the case of strong mutation provides an interesting topic for further study.

3.5 Conclusion

The puzzle of the divergence between the predictions of the replicator dynamics and the Moran process finds its resolution in identifying a shortcoming of stochastic stability analysis. The cause of mis-prediction by stochastic stability is the assumption of vanishing mutation. Polymorphism, which cannot be stochastically stable, can be the most probable long run outcome of the Moran process.

I have shown that, under a range of values of population size, mutation rate, and intensity of selection, the Moran process leads to polymorphisms which dominate the long run behavior of the process. My results show that anti-coordination games, and games containing anti-coordination subgames, can exhibit this behavior for a broad range of conditions. For the 2×2 anti-coordination games considered here, 'strong mutation' provides a sufficient condition for mis-prediction by stochastic stability analysis of the long run behavior of the Moran process. Moreover, in the presence of strong mutation, the Moran process will typically spend most of its time near the specific polymorphic state that is asymptotically stable under replicator dynamics.

We have also seen that strong mutation will be satisfied by a range of real-world evolutionary processes. This is particularly true when population sizes are large, such as with bacterial colonies, and when mutation or noise rates are high, as is typical in the transmission of behavior in models of cultural evolution. In such cases, we can anticipate that the behavior of the Moran process will be mis-characterized by stochastic stability, and will realign with the predictions of the replicator dynamics.

The upshots of our analysis are that we can characterize the conditions under which an evolutionary process described by the Moran process (1) will sustain long run diversity, (2) realign with the predictions of the replicator dynamics, and (3) should not be analyzed using stochastic stability. Our moral is that, when we anticipate attracting polymorphic equilibria—that is, when a population interaction structure is characterized by anti-coordination—stochastic stability may be an unreliable predictor of even the long term behavior of evolution. In such cases, analysis should proceed by computing the stationary distribution explicitly using representative values of population size, mutation rate, and intensity of selection. When such an approach is not feasible, simulation methods must suffice. In mathematical modeling, we must attend to idealizations not only in the models themselves but also within the techniques with which those models are analyzed.

3.6 Mathematical Appendix

For the following proofs, we consider a game under the Moran process with population size $N \in \mathbb{N}$, mutation rate $\eta \in (0, 1/2)$, and intensity of selection $w \in [0, 1]$, characterized by any 2×2 payoff matrix $A = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ where $a, b, c, d > 0$,[22] payoff functions f_i and g_i, and transition matrix $P_{i,j}$ over the finite state space $\chi = \{0, 1, \ldots, N\}$. This yields the ergodic process $\{X_t^{N,\eta,w}\}$.

Let the fitness of each type at a particular intensity of selection be denoted $f^w \equiv f|_w$, $g^w \equiv g|_w$. Similarly, for transition probabilities, $P^w \equiv P|_w$, and stationary distributions $\mu^w \equiv \mu|_w$. We will omit the state subscript i, when there is no risk of confusion.

Proof of Lemma 1. We want to show that, in the absence of selection, the peak of the stationary distribution is a polymorphic state if, and only if, strong

[22] Note that the stipulation of positive payoffs is required as positive fitness values are needed for the Moran process to be well-defined.

mutation $\eta(N+2) > 1$ holds.

Since $\{X_t^{N,\eta,w}\}$ is an ergodic process when $\eta > 0$, we are guaranteed that it has a unique stationary distribution $\mu = \langle \mu_0, \ldots, \mu_N \rangle$. Further, since we are considering a 2-strategy game, we know that μ satisfies the detailed balance condition $\mu_i P_{i,i-1} = \mu_{i-1} P_{i-1,i}$ (Sandholm, 2009, Ch.12). From this, it follows that $\mu_i > \mu_{i-1}$ just in case $P_{i,i-1} < P_{i-1,i}$. That is, a state i has greater mass in the stationary distribution than its preceding state $i-1$ just in case the transition probability from i to $i-1$ is less than the transition probability from $i-1$ to i.

Set intensity of selection to zero $w = 0$, giving $f^0 = g^0 = 1$. From these fitnesses we determine the relevant transition probabilities.

$$\begin{aligned}
P_{i,i-1} &= \frac{i\,(N-i+\eta(2i-N))}{N^2}\\
P_{i,i+1} &= \frac{(N-i)\,(i+\eta(N-2i))}{N^2}\\
P_{i-1,i} &= \frac{(N-i-1)\,(i-1+\eta(N-2i-2))}{N^2}
\end{aligned}$$

Now, we find the conditions under which $P_{i,i-1} < P_{i-1,i}$ and $P_{i,i-1} > P_{i-1,i}$ in terms of i, N, and η. This will tell us when the mass of states in the stationary distribution is increasing, and when it is decreasing. Unpacking the inequality $P_{i,i-1} < P_{i-1,i}$, we get

$$\frac{i\,(N-i+\eta(2i-N))}{N^2} < \frac{(N-i-1)\,(i-1+\eta(N-2i-2))}{N^2}$$

which, with some algebra, yields

$$(1-\eta(N+2))\,(N-2i+1) < 0. \tag{$*$}$$

We make the necessary restrictions, $2 \leq N$ and $1 \leq i \leq N$, and denote the term on the left hand side of the inequality $(*)$ by h. We see that, when $\eta(N+2) > 1$, $i < \frac{N+1}{2}$ implies $h < 0$ and $i > \frac{N+1}{2}$ implies $h > 0$. Whereas, when $\eta(N+2) < 1$, $i < \frac{N+1}{2}$ implies $h > 0$ and $i > \frac{N+1}{2}$ implies $h < 0$.

That is, when strong mutation obtains, the mass μ_i of a state i in the stationary distribution is greater than that of its preceding state $i-1$ over the first half of the state space $i < \frac{N+1}{2}$, and less than that of its preceding state over the second half of the state space $i > \frac{N+1}{2}$. Thus, the stationary distribution μ must exhibit a unique mode exactly at (or, when N is even, at the states directly adjacent to) the center of the state space $i = \frac{N+1}{2}$. And when strong mutation does not obtain, the relation between the mass of adjacent states are precisely reversed, and the stationary distribution must exhibit two modes, one at each of the monomorphic states $i = 0$ and $i = N$.

Thus, in the absence of selection, strong mutation is necessary and sufficient for the mode of the stationary distribution to be a polymorphic state, and any polymorphic mode will be at the midpoint of the state space. □

To tackle our theorem, first we prove some helpful lemmas.

Lemma 2. *All else being equal, increasing intensity of selection exaggerates selection in favor of the fitter type. That is, if $w < w'$, then $f^w > g^w$ just in case*

$$\frac{if^w}{if^w + (N-i)g^w} < \frac{if^{w'}}{if^{w'} + (N-i)g^{w'}} \quad \textit{and} \quad \frac{(N-i)g^w}{if^w + (N-i)g^w} > \frac{(N-i)g^{w'}}{if^{w'} + (N-i)g^{w'}}.$$

Proof. Consider two versions of the same process, $\{X_t^{N,\eta,w}\}$ and $\{X_t^{N,\eta,w'}\}$, differing only in that the latter has greater intensity of selection, $w < w'$. Suppose $f_i^w > g_i^w$ for some $i \in \chi$. Then $f^w - g^w = wk$ and $f^{w'} - g^{w'} = w'k$ where $k = (a(i-1) + (b(N-i))) - (ci + d(N-i-1))/(N-1)$. Hence $\frac{f^w - g^w}{f^{w'} - g^{w'}} = \frac{wk}{w'k} = \frac{w}{w'} < 1$. We now have that $0 < f^w - g^w < f^{w'} - g^{w'}$, and so $\frac{f^w}{g^w} < \frac{f^{w'}}{g^{w'}}$. We turn to the selection terms of our transition probabilities, and observe that the following inequalities are equivalent.

$$\frac{f^w}{g^w} < \frac{f^{w'}}{g^{w'}}$$

$$\frac{i}{N-i}\frac{f^w}{g^w} < \frac{i}{N-i}\frac{f^{w'}}{g^{w'}}$$

$$if^w(N-i)g^{w'} < if^{w'}(N-i)g^w$$

$$if^w(N-i)g^{w'} + (if^w \cdot if^{w'}) < if^{w'}(N-i)g^w + (if^w \cdot if^{w'})$$

$$if^w((N-i)g^{w'} + if^{w'}) < if^{w'}((N-i)g^w + if^w)$$

$$\frac{if^w}{if^w + (N-i)g^w} < \frac{if^{w'}}{if^{w'} + (N-i)g^{w'}}.$$

By similar reasoning, the following are equivalent

$$\frac{f^w}{g^w} < \frac{f^{w'}}{g^{w'}}$$

$$\frac{(N-i)g^{w'}}{if^{w'} + (N-i)g^{w'}} < \frac{(N-i)g^w}{if^w + (N-i)g^w},$$

as required. □

Lemma 3. *All else being equal, increasing intensity of selection exaggerates transition probabilities in favor of the fitter type. That is, if $w < w'$, then $f^w > g^w$ just in case*

$$P^w_{i,i+1} < P^{w'}_{i,i+1} \quad \textit{and} \quad P^w_{i,i-1} > P^{w'}_{i,i-1}.$$

Proof. Denote $A \equiv \frac{if^w}{if^w+(N-i)g^w}$, $A' \equiv \frac{if^{w'}}{if^{w'}+(N-i)g^{w'}}$, $B \equiv \frac{(N-i)g^w}{if^w+(N-i)g^w}$, and $B' \equiv \frac{(N-i)g^{w'}}{if^{w'}+(N-i)g^{w'}}$. Suppose $w < w'$, and $f^w_i > g^w_i$ for some $i \in \chi$. From Lemma 2, we have that $f^w > g^w$ just in case $A < A'$ and $B > B'$. We will make use of the fact that $B = 1 - A$ and $B' = 1 - A'$. Let $\eta < 1/2$. Then the following inequalities are equivalent.

$$\begin{gathered} A < A' \\ A(1-2\eta) + \eta < A'(1-2\eta) + \eta \\ (1-\eta)A + \eta(1-A) < (1-\eta)A' + \eta(1-A') \\ (1-\eta)A + \eta B < (1-\eta)A' + \eta B' \\ (1-\eta)\frac{N-i}{N}A + \eta\frac{N-i}{N}B < (1-\eta)\frac{N-i}{N}A' + \eta\frac{N-i}{N}B' \\ P^w_{i,i+1} < P^{w'}_{i,i+1}. \end{gathered}$$

By similar reasoning, the following inequalities are equivalent.

$$\begin{gathered} B > B' \\ (1-\eta)\frac{i}{N}B + \eta\frac{i}{N}A > (1-\eta)\frac{i}{N}B' + \eta\frac{i}{N}A' \\ P^w_{i,i-1} > P^{w'}_{i,i-1}, \end{gathered}$$

as required. □

Proof of Theorem 1. Let all else be as before, except our 2 × 2 symmetric game is now characterized by anti-coordination payoffs $a < c$, $b > d$, with the extra condition required for finite games that $b - d > \frac{a-d}{N} > a - c$.

From Lemma 1, we have that, in the absence of selection $w = 0$, strong mutation $\eta(N+2) > 1$ is necessary and sufficient for $\mu_{i-1} < \mu_i$ for $i \leq \lfloor \frac{N+1}{2} \rfloor$ and $\mu_{i-1} > \mu_i$ for $i > \lceil \frac{N+1}{2} \rceil$.

For nonzero intensity of selection $w > 0$, we will show that $f^w > f^0$ and $g^w < g^0$ for some range of states before a polymorphic critical point i^*. As we will show, it follows that $P^w_{i-1,i} > P^0_{i-1,i}$ and $P^w_{i,i-1} < P^0_{i,i-1}$ which in turn implies that $\frac{P^w_{i-1,i}}{P^w_{i,i-1}} > \frac{P^0_{i-1,i}}{P^0_{i,i-1}} > 1$. From the detailed balance conditions, this yields $\mu_{i-1} < \mu_i$ when $i \leq \lfloor i^* \rfloor$. Similarly, we will find that $\frac{P^w_{i-1,i}}{P^w_{i,i-1}} < \frac{P^0_{i-1,i}}{P^0_{i,i-1}} < 1$, and hence $\mu_{i-1} < \mu_i$ after the critical point, when $i > \lceil i^* \rceil$. This will conclude the proof.

Let $w > 0$. First, we find our critical point i^*. Recall the fitness functions for each type

$$f^w_i = 1 - w + w\frac{a(i-1) + b(N-i)}{N-1} \quad \text{and} \quad g^w_i = 1 - w + w\frac{ci + d(N-i-1)}{N-1}.$$

To find the critical point, we solve for when each type is fitter than the other.

$$\begin{gathered} f^w - g^w > 0 \\ a(i-1) + b(N-i) - ci - d(N-i-1) > 0 \\ i(a-b-c+d) + N(b-d) + (d-a) > 0 \\ i < \frac{N(b-d) + (d-a)}{b+c-a-d} \end{gathered}$$

Hence, $f^w > g^w$ just in case $i < i^* = \frac{N(b-d)+(d-a)}{b+c-a-d}$. Note that we can confirm that our interior critical point i^* is indeed well-defined as $b - d > \frac{a-d}{N} > a - c$ implies that $0 < \frac{N(b-d)+(d-a)}{b+c-a-d} < N$.

From Lemma 1, we know that, whenever the strong mutation condition is satisfied, the mass of stationary distribution of the process in the absence of selection μ^0_i is increasing over the first half of the state space $i < \frac{N+1}{2}$, and decreasing over the second half $i > \frac{N+1}{2}$.

From Lemma 3, it follows from $w > 0$ that $f^w > g^w$ implies $P^0_{i,i+1} < P^w_{i,i+1}$ and $P^0_{i,i-1} < P^w_{i,i-1}$, which obtains for $i < i^*$, and $f^w < g^w$ implies $P^0_{i,i+1} > P^w_{i,i+1}$ and $P^0_{i,i-1} > P^w_{i,i-1}$, which obtains for $i > i^*$. Hence, when $f^w > g^w$ and $\mu^0_i > \mu^0_{i-1}$, we have that $\frac{P^w_{i-1,i}}{P^w_{i,i-1}} > \frac{P^0_{i-1,i}}{P^0_{i,i-1}} > 1$. And, when $f^w < g^w$ and $\mu^0_i < \mu^0_{i-1}$, we have that $\frac{P^w_{i-1,i}}{P^w_{i,i-1}} < \frac{P^0_{i-1,i}}{P^0_{i,i-1}} < 1$.

From this, and the detailed balance conditions, we know that μ^w must be increasing for $i \leq \min\{\lfloor i^* \rfloor, \lfloor \frac{N+1}{2} \rfloor\}$, and decreasing for $i > \max\{\lceil i^* \rceil, \lceil \frac{N+1}{2} \rceil\}$. Thus, we have that the stationary distribution μ^w must find it maximum value at a polymorphic state somewhere in a state between $i^* = \frac{b-d+(\frac{d-a}{N})}{b+c-a-d}$ and $\frac{N+1}{2}$. □

Proof of Corollary 1. Consider an anti-coordination game under the Moran process, as before. Suppose the strong mutation condition $\eta(N+2) > 1$ is satisfied, and consider two intensities of selection w, w' where $0 < w < w' \leq 1$. Then, for every population state prior to the critical point $i^* = \frac{d-b+(\frac{a-d}{N})}{d-c-b+a}$

we know that $f^w > g^w$ and $f^{w'} > g^{w'}$. By lemma 3, $w < w'$ implies that $P^w_{i,i-1} > P^{w'}_{i,i-1}$ and $P^w_{i-1,i} < P^{w'}_{i-1,i}$. So $\frac{P_{i,i-1}}{P^w_{i-1,i}} > \frac{P^{w'}_{i,i-1}}{P^{w'}_{i-1,i}}$ which gives us, from the detailed balance conditions, that $\frac{\mu^w_i}{\mu^w_{i-1}} < \frac{\mu^{w'}_i}{\mu^{w'}_{i-1}}$.

This means that the increase in mass $(\mu_i - \mu_{i-1})$ in every state states prior to the critical point i^* is greater (though, of course, it may be still be negative for some states between i^* and $\frac{N+1}{2}$) for $\mu^{w'}$ than for μ^w. It is easy to see that the inverse inequalities obtain for states after the critical point i^*, and so the rate of decrease in mass is greater for $\mu^{w'}$ than for μ^w for $i > i^*$.

By the conservation of mass of the stationary distribution, $\sum_i \mu_i = 1$, if the rate of increase (of mass) for every state of a distribution $\mu^{w'}$ is greater than another μ^w to the left of a critical point $i < i^*$ and the rate of decrease for every state of $\mu^{w'}$ is greater than for μ^w to the right of that critical point $i^* > i$, then $\mu^{w'}$ must place greater mass than μ^w on the state(s) nearest the critical point. Hence, the mass of the state(s) nearest the critical point $i^* = \frac{N(b-d)+(d-a)}{b+c-a-d}$ is increasing in intensity of selection. □

3.7 Computational Appendix

A GUI for exploring the dynamics of the frequency-dependent Moran process under 2 × 2 strategic games is available at: `https://amohseni.shinyapps.io/Moran-Process/`.

The full R source code for all simulations can be found at:
`https://github.com/amohseni/Frequency-Dependent-Moran-Process`

Bibliography

S. E. Asch. Opinions and Social Pressure. *Scientific American*, 193(5):31–35, 1955.

A. V. Banerjee. A Simple Model of Herd Behavior. *The Quarterly Journal of Economics*, 107 (3):797–817, 1992.

E. C. Barnes. The roots of predictivism. *Studies in History and Philosophy of Science Part A*, 45:46–53, 2014.

J. Barrett, B. Skyrms, and A. Mohseni. Self-Assembling Networks. *British Journal for the Philosophy of Science*, (70):301–325, 2017.

B. Baumgaertner, B. Devezer, E. O. Buzbas, and L. G. Nardin. Openness and reproducibility: Insights from a model-centric approach. *PLos ONE*, 14(5):e0216125, 2019.

M. Benaïm and J. Weibull. Deterministic Approximation of Stochastic Evolution in Games. *Econometrica*, 71(3):873–903, 2003.

M. Benaïm and J. Weibull. Mean-field approximation of stochastic population processes in games. Technical Report 1979, 2009.

D. J. Benjamin, J. O. Berger, M. Johannesson, B. A. Nosek, E. J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, D. Cesarini, C. D. Chambers, M. Clyde, T. D. Cook, P. De Boeck, Z. Dienes, A. Dreber, K. Easwaran, C. Efferson, E. Fehr, F. Fidler, A. P. Field, M. Forster, E. I. George, R. Gonzalez, S. Goodman, E. Green, D. P. Green, A. G. Greenwald, J. D. Hadfield, L. V. Hedges, L. Held, T. Hua Ho, H. Hoijtink, D. J. Hruschka, K. Imai, G. Imbens, J. P. Ioannidis, M. Jeon, J. H. Jones, M. Kirchler, D. Laibson, J. List, R. Little, A. Lupia, E. Machery, S. E. Maxwell, M. McCarthy, D. A. Moore, S. L. Morgan, M. Munafó, S. Nakagawa, B. Nyhan, T. H. Parker, L. Pericchi, M. Perugini, J. Rouder, J. Rousseau, V. Savalei, F. D. Schönbrodt, T. Sellke, B. Sinclair, D. Tingley, T. Van Zandt, S. Vazire, D. J. Watts, C. Winship, R. L. Wolpert, Y. Xie, C. Young, J. Zinman, and V. E. Johnson. Redefine statistical significance. *Nature Human Behaviour*, 2:6–10, 2018.

Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

S. Bikhchandani, D. Hirshleifer, and I. Welch. A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades. *Journal of Political Economy*, 1992.

K. Binmore and L. Samuelson. Muddling Through: Noisy Equilibrium Selection. *Journal of Economic Theory*, 74(2):235–265, 1997.

K. Binmore, L. Samuelson, and R. Vaughan. Musical chairs: modeling noisy evolution. *Games and economic behavior*, 11(1):1–35, 1995.

A. Bird. Understanding the Replication Crisis as a Base Rate Fallacy. *The British Journal for the Philosophy of Science*, 0:1–31, 2020.

D. Bishop. Rein in the four horsemen of irreproducibility. *Nature*, 568:435, 2019.

R. Bond and P. B. Smith. Culture and conformity: A meta-analysis of studies using asch's (1952b, 1956) line judgment task. *Psychological Bulletin*, 119(1):111–137, 1996.

S. Bowles and H. Gintis. *A Cooperative Species: Human Reciprocity and its Evolution.* 2011.

R. Boyd and P. J. Richerson. *Culture and the Evolutionary Process*, volume 175. 1985.

L. K. Bright. On fraud. *Philosophical Studies*, 174:291–310, 2017.

J. Bruner and C. O'Connor. Power, Bargaining , and Collaboration. In *Scientific Collaboration and Collective Knowledge*, pages 1–25. 2016.

J. P. Bruner and B. Holman. Self-correction in science: Meta-analysis, bias and social structure. *Studies in history and philosophy of science*, 78:93 – 97, 2019.

R. Cressman. *Evolutionary Dynamics and Extensive Form Games.* MIT Press, 2003.

R. Cressman and Y. Tao. The replicator equation and other game dynamics. *Proceedings of the National Academy of Sciences*, 111(Supplement_3):10810–10817, 2014.

B. Devezer, L. G. Nardin, B. Baumgaertner, and E. O. Buzbas. Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PLOS ONE*, 14:1–23, 05 2019.

H. Douglas and P. D. Magnus. State of the Field: Why novel prediction matters. *Studies in History and Philosophy of Science Part A*, 44:580–589, 2013.

J. W. Drake, B. Charlesworth, D. Charlesworth, and J. F. Crow. Rates of spontaneous mutation. *Genetics*, 148(4):1667–1686, 1998.

A. Dreber, T. Pfeiffer, J. Almenberg, S. Isaksson, B. Wilson, Y. Chen, B. A. Nosek, M. Johannesson, and K. W. Wachter. Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences of the United States of America*, 112(50):15343–15347, 2015.

A. Etz and J. Vandekerckhove. A Bayesian perspective on the reproducibility project: Psychology. *PLoS ONE*, 2016.

A. Franklin. *Shifting standards: Experiments in particle physics in the twentieth century.* University of Pittsburgh Press, 2013.

D. Fudenberg and L. Imhof. Stochastic Evolution as a Generalized Moran Process. *Unpublished Manuscript*, pages 1–26, 2004.

D. Fudenberg and L. A. Imhof. Imitation processes with small mutations. *Journal of Economic Theory*, 131(1):251–262, 2006.

D. Fudenberg, M. A. Nowak, C. Taylor, and L. A. Imhof. Evolutionary game dynamics in finite populations with strong selection and weak mutation. *Theoretical Population Biology*, 70(3):352–363, 2006.

J. García and A. Traulsen. The structure of mutations and the evolution of cooperation. *PLoS ONE*, 7(4), 2012.

J. K. Goeree, A. Riedl, and A. Ule. In search of stars: Network formation among heterogeneous agents. *Games and Economic Behavior*, 67(2):445–466, 2009.

I. J. Good. On the principle of total evidence. *British Journal for the Philosophy of Science*, 17(4):319–321, 1967.

S. Goyal. Princeton University Press, 2007.

P. Grim, D. J. Singer, S. Fisher, A. Bramson, W. J. Berger, C. Reade, C. Flocken, and A. Sales. Scientific networks on data landscapes: Question difficulty, epistemic success, and convergence. *Episteme*, 2013.

W. D. Hamilton. Extraordinary Sex Ratios. *Science*, 156(3774):477–488, 1967.

M. Harper and D. Fryer. Stationary stability for evolutionary dynamics in finite populations. *Entropy*, 18(9), 2016.

M. L. Head, L. Holman, R. Lanfear, A. T. Kahn, and M. D. Jennions. The Extent and Consequences of P-Hacking in Science. *PLoS Biology*, 13(3):e1002106, 2015.

R. Heesen. Communism and the Incentive to Share in Science. *Philosophy of Science*, 84(4): 698–716, 2017.

R. Heesen. Why the reward structure of science makes reproducibility problems inevitable. *Journal of Philosophy*, 115(12):661–674, 2018.

C. Hitchcock and E. Sober. Prediction versus accommodation and the risk of overfitting. *British Journal for the Philosophy of Science*, 55(1):1–34, 2004.

J. R. Hollenbeck and P. M. Wright. Harking, Sharking, and Tharking: Making the Case for Post Hoc Analysis of Scientific Data. *Journal of Management*, 43(1):5–18, 2016.

L. A. Imhof and M. A. Nowak. Evolutionary game dynamics in a Wright-Fisher process. *Journal of Mathematical Biology*, 52(5):667–681, 2006.

J. P. Ioannidis. Why most published research findings are false. *PLOS Medicine*, 2(8):e124, 2005.

J. P. Ioannidis. The proposal to lower P value thresholds to .005. *Journal of the American Medical Association*, 319(14):1429–1430, 2018.

L. K. John, G. Loewenstein, and D. Prelec. Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, 23(5):524–532, 2012.

S. Karlin and H. E. Taylor. *A First Course in Stochastic Processes: Second Edition*. 2012.

N. L. Kerr. HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3):196–217, 1998.

P. Kitcher. The Division of Cognitive Labor. *The Journal of Philosophy*, 87(1):5, 1990.

P. Kitcher. *The Advancement of Science: Science Without Legend, Objectivity Without Illusions*. Oxford University Press, 1993.

S. Kumar and S. Subramanian. Mutation rates in mammalian genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 99(2):803–808, 2002.

H. J. Lamberink, W. M. Otte, M. R. Sinke, D. Lakens, P. P. Glasziou, J. K. Tijdink, and C. H. Vinkers. Statistical power of clinical trials increased while effect size remained stable: an empirical analysis of 136,212 clinical trials between 1975 and 2014. *Journal of Clinical Epidemiology*, 102:123 – 128, 2018.

H. E. Landemore. Why the Many Are Smarter than the Few and Why It Matters. *Journal of Public Deliberation*, 1(8), 2012.

E. L. Lehmann. The fisher, Neyman–Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association*, 88:1242–1249, 1993.

K. Leung. Presenting Post Hoc Hypotheses as A Priori: Ethical and Theoretical Issues. *Management and Organization Review*, 7:471–479, 2011.

L. Lyons. Discovering the Significance of 5 Sigma. *arXiv*, 2013. URL `https://arxiv.org/abs/1310.1284`.

L. Lyons. Statistical Issues in Searches for New Physics. *arXiv*, 2015. URL `https://arxiv.org/abs/1409.1903`.

E. Machery. What is a replication? *Philosophy of Science*, 87(4):545–567, 2020.

Z. Maniadis, F. Tufano, and J. A. List. One swallow doesn't make a summer: New evidence on anchoring effects. *American Economic Review*, 104(1):277–90, 2014.

J. Maynard Smith. The theory of games and the evolution of animal conflicts. *Journal of Theoretical Biology*, 47(1):209–221, 1974.

D. Mayo. *Error and the Growth of Experimental Knowledge*. University of Chicago Press, 2006.

D. Mayo. Some surprising facts about (the problem of) surprising facts. *Studies in History and Philosophy of Science Part A*, 45:79 – 86, 2014.

D. Mayo. *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge University Press, 2019.

D. G. Mayo and A. Spanos. Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. *British Journal for the Philosophy of Science*, 57:323–357, 2006.

D. G. Mayo and A. Spanos. Error Statistics. In *Philosophy of Statistics*, volume 7, page 153–198. 2011.

C. Mayo-Wilson and S. C. Fletcher. Evidence in Classical Statistics. In M. Lasonen-Aarnio and C. Littlejohn, editors, *Routledge Handbook of the Philosophy of Evidence*. Routledge, 2019.

C. Mayo-Wilson, K. J. S. Zollman, and D. Danks. Wisdom of crowds versus groupthink: Learning in groups and in isolation. *International Journal of Game Theory*, 42(3):695–723, 2013.

H. Mercier and H. Landemore. Reasoning is for arguing: Understanding the successes and failures of deliberation. *Political Psychology*, 2012.

J. S. Mill. *A System of Logic, Ratiocinative and Inductive*. Routlege, 1843.

A. Mohseni. The Limitations of Equilibrium Concepts in Evolutionary Games. *Unpublished Manuscript*, 2017.

P. A. Moran. *The Statistical Processes of Evolutionary Theory*. Clarendon Press, Oxford, 1962.

T. J. Morganand and K. N. Laland. The biological bases of conformity, 2012.

K. R. Murphy and H. Aguinis. HARKing: How badly can cherry-picking and question trolling produce bias in published results? *Journal of Business and Psychology*, 34:1–17, 2019.

M. Novak. *Evolutionary Dynamics: Exploring the Equations of Life*. Number 3. 2007.

M. A. Nowak, A. Sasaki, C. Taylor, and D. Fudenberg. Emergence of cooperation and evolutionary stability in finite populations. *Nature*, 428(6983):646–650, 2004.

C. O'Connor, L. Bright, and J. Bruner. The Emergence of Intersectional Disadvantage. *Social Epistemology*, 1(33):23–41.

H. Ohtsuki, P. Bordalo, and M. A. Nowak. The one-third law of evolutionary dynamics. *Journal of Theoretical Biology*, 249(2):289–295, 2007.

O. A. Panagiotou, J. P. Ioannidis, J. N. Hirschhorn, G. R. Abecasis, T. M. Frayling, M. I. McCarthy, C. M. Lindgren, T. H. Beaty, N. Eriksson, C. Polychronakos, S. Kathirensan, R. M. Plenge, R. Spritz, H. Payami, E. R. Martin, J. Vance, W. H. Su, Y. S. Chang, J. X. Bei, Y. X. Zeng, G. Paré, S. V. Faraone, B. Neale, R. J. Anney, B. J. Traynor, A. Scherag, J. Hebebrand, A. Hinney, P. Froguel, D. Meyre, S. J. Chanock, and W. Kesheng. What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *International Journal of Epidemiology*, 41(1):273–86, 2012.

S. Pawel and L. Held. Probabilistic forecasting of replication studies. *PLoS ONE*, 4(15): e0231416, 2020.

K. Popper. *The logic of scientific discovery.* Routlege, 1934.

F. Romero. Philosophy of science and the replicability crisis. *Philosophy Compass*, 14:e12633, 2019.

F. Romero. The Division of Replication Labor. *Philosophy of Science*, 87(5):1014–1025, 2020.

F. Romero and J. Sprenger. Scientific self-correction: the Bayesian way. *Synthese*, 2020.

S. Rosenstock, J. Bruner, and C. O'Connor. In Epistemic Networks, Is Less Really More? *Philosophy of Science*, 84(2):234–252, 2017.

M. Rubin. When does HARKing hurt? Identifying when different types of undisclosed post hoc hypothesizing harm scientific progress. *Review of General Psychology*, 21(4):308–320, 2017.

M. Rubin. The Costs of HARKing. *The British Journal for the Philosophy of Science*, page 1–30, 2019.

W. H. Sandholm. Simple formulas for stationary distributions and stochastically stable states. *Games and Economic Behavior*, 59(1):154–162, 2007.

W. H. Sandholm. Population Games and Evolutionary Dynamics. *Population (English Edition)*, pages xvi, 556, 2009.

W. H. Sandholm. Orders of limits for stationary distributions, stochastic dominance, and stochastic stability. *Theoretical Economics*, 5(1):1–26, 2010.

W. H. Sandholm. Stochastic imitative game dynamics with committed agents. *Journal of Economic Theory*, 147(5):2056–2071, 2012.

L. Savage. *The Foundations of Statistics.* Wiley Publications in Statistics, 1954.

G. Schurz. Bayesian pseudo-confirmation, use-novelty, and genuine confirmation. *Studies in History and Philosophy of Science Part A*, 45:87 – 96, 2014.

B. Skyrms. *The Dynamics of Rational Deliberation.* Harvard University Press, 1990.

B. Skyrms. *Evolution of the social contract.* Cambridge University Press, 1996.

L. Smith and P. Sørensen. Pathological outcomes of observational learning. *Econometrica*, 68 (2):371–398, 2000.

M. Strevens. The role of the priority rule in science. *The Journal of Philosophy*, 100:55–79, 2003.

M. Strevens. Herding and the quest for credit. *Journal of Economic Methodology*, 20(1):19–34, 2013.

D. Szucs and J. P. Ioannidis. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology*, 3(15):e2000797, 2017.

C. Taylor, D. Fudenberg, A. Sasaki, and M. A. Nowak. Evolutionary game dynamics in finite populations. *Bulletin of Mathematical Biology*, 66(6):1621–1644, 2004.

P. D. Taylor and L. B. Jonker. Evolutionarily stable strategies and game dynamics. *Mathematical Biosciences*, 40(1-2):145–156, 1978.

O. Tenaillon, J. E. Barrick, N. Ribeck, D. E. Deatherage, J. L. Blanchard, A. Dasgupta, G. C. Wu, S. Wielgoss, S. Cruveiller, C. Médigue, D. Schneider, and R. E. Lenski. Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature*, 536(7615):165–170, 2016.

P. E. Tetlock. *Expert political judgment: How good is it? How can we know?, new edition*. Princeton University Press, 2017.

A. Traulsen and C. Hauert. Stochastic Evolutionary Game Dynamics. In *Reviews of Nonlinear Dynamics and Complexity*, volume 2, pages 25–61. 2010.

D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393 (6684):440–442, 1998.

J. Worrall. Prediction and accommodation revisited. *Studies in History and Philosophy of Science Part A*, 45:54–61, 2014.

H. P. Young. The Evolution of Conventions. *Econometrica*, 61(1):57–84, 1993.

H. P. Young. *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Number 461. Princeton Univeristy Press, 1998.

H. P. Young. The Diffusion of Innovations in Social Networks. *The Economy as an Evolving Complex System*, III(1966):267–282, 2005.

H. P. Young. The Evolution of Social Norms. *Annual Review of Economics*, 7(1):359–387, 2015.

K. J. S. Zollman. The Communication Structure of Epistemic Communities. *Philosophy of Science*, 74(5):574–587, 2007.

K. J. S. Zollman. The Epistemic Benefit of Transient Diversity. *Erkenntnis*, 72(1):17–35, 2009.

K. J. S. Zollman. Social structure and the effects of conformity. *Synthese*, pages 317–340, 2010.

K. J. S. Zollman. Network epistemology: Communication in epistemic communities. *Philosophy Compass*, 8(1):15–27, 2013.