# Limitations on Robust Ratings and Predictions

Tim Muller(✉), Yang Liu, and Jie Zhang

Nanyang Technological University, Singapore, Singapore
t.j.c.muller@gmail.com

**Abstract.** Predictions are a well-studied form of ratings. Their objective nature allows a rigourous analysis. A problem is that there are attacks on prediction systems and rating systems. These attacks decrease the usefulness of the predictions. Attackers may ignore the incentives in the system, so we may not rely on these to protect ourselves. The user must block attackers, ideally before the attackers introduce too much misinformation. We formally axiomatically define robustness as the property that no rater can introduce too much misinformation. We formally prove that notions of robustness come at the expense of other desirable properties, such as the lack of bias or effectiveness. We also show that there do exist trade-offs between the different properties, allowing a prediction system with limited robustness, limited bias and limited effectiveness.

## 1 Introduction

Ratings are an important tool in online cooperation. Ratings are used in, e.g., recommender systems, trust and reputation systems, e-commerce systems and security systems [10–12]. We reason about a specific type of predictions, namely those that we can judge in hindsight – called predictions. Prediction are also an interesting topic of research in themselves [1]. Typically, users that give predictions that are better (accurate or honest) are rewarded by becoming more credible. However, there are incentives outside of the system that may drive a user to give worse (inaccurate or dishonest) predictions. These unfair ratings attacks are well-known in literature, and found to occur in real systems. On a robust prediction system, the impact of these unfair ratings is limited.

A standard technique in prediction systems is to have a mechanism to encourage users to behave in a certain way, by setting the right incentives. However, in practice, users may have a bigger incentive to give bad predictions. We know that users attack systems by providing false predictions [9,14] despite losing credit within the system. A user that ignores the incentives of a system is called an attacker. In other words, an attacker does not necessarily care about the rewards and punishments that the prediction system sets, because incentives outside of the system (e.g. bribes, collusion) are greater.

As we cannot modify the behaviour of these attackers, we must resort to interpreting the predictions in a robust fashion. Specifically, we must somehow limit the impact of unfair ratings. In this paper, we introduce notions of robustness (differing in strength) that codify that the amount of noise that a single


S.M. Habib et al. (Eds.): IFIPTM 2016, IFIP AICT 473, pp. 113–126, 2016.
DOI: 10.1007/978-3-319-41354-9_8

agent can introduce is limited. We have a threefold motivation for the exact formulations: intuitive grounds, information theory and hypothesis testing. Given our definition of robustness, we can prove a specific prediction system to be robust.

Robustness comes at a cost. With no tolerance towards misinformation, any useful way of using predictions is impossible. For weaker robustness requirements, we have more subtle impossibilities regarding the use of predictions. One main contribution in this paper is a general and rigourous proof that robustness, bias and effectiveness are tradeoffs, and that certain combinations are impossible. The proofs are axiomatic, meaning that we have axioms for the various levels of robustness, bias and effectiveness, and we prove that no model can satisfy all of them. Specifically:

- No meaningful model exists for absolute robustness (no tolerance towards misinformation).
- Any model for strict robustness (fixed misinformation threshold) has some bias and a finite lifespan.
- Any model for weak robustness (growing tolerance towards misinformation) has some bias and cannot be fully effective.

The results are summarised in Table 1.

Fortunately, if we are willing to make the trade-offs, then robust models do exist. We show that a prediction system with strict robustness can exist and be useful despite its hampered effectiveness. Similarly, we also show that a prediction system with weak robustness can be implemented and be far more effective. These results extend only to prediction systems, since they rely on the user knowing the validity of the predictions after the fact.

The paper is organised as follows. First we discuss work related to our domain, but also impossibility results in social choice which inspired the methodology. In Sect. 3, we present the requirements that we want a prediction system to fulfill, in natural language. Then, in Sect. 4, we present a formal model of predictions, events, filters and misinformation. In Sect. 5, we formalise the requirements in that model. In Sect. 6, we establish the relationships between the axioms – particularly we close the bridge between the information-theoretic and the statistical perspective on the quality of predictions. In Sect. 7, we prove the impossibility results. All the limitations of robust prediction systems can be found in this section. In Sect. 8, we prove the existence of prediction systems that have robustness, albeit with considerable reduction of effectiveness. In these latter two sections, non-technical formulations of the results are presented in bold font. We provide a conclusion in Sect. 9.

## 2 Related Work

Our original research interest lies in robust ratings, as e.g. in [16,17]. There, the ratings are quantified using information theory. This idea is not novel, as e.g. [11] also uses information theory to quantify ratings. Our novelty is that we

**Table 1.** Effectiveness given levels of robustness (**AR**, **SR**, **WR**), bias (**SU**, **WU**) and non-prescience (**T**).

| | WU | SU | WU & T | SU & T | T |
|---|---|---|---|---|---|
| **AR** | 0 | 0 | 0 | 0 | 0 |
| **SR**($\theta$) | $2^{\theta}$ | 0 | $\theta$ | 0 | $\theta$ |
| **WR**($f$) | $> 2^{f(1)}$ | 0 | $f(n)$ | 0 | $f(n)$ |

are able to formulate a system with a strict robustness cutoff. The damage of attacks is strictly limited. However, for the system to work, the ratings must be predictions – verifiable in hindsight.

Prediction systems are widely used and studied [3,8]. An important type of prediction system is a prediction market. There has been lots of research on prediction markets, especially their resistance against manipulation [4,6,9]. An inherent problem of prediction markets is that raters insensitive to the system's incentives have absolute freedom to manipulate [4,6]. Our approach limits the influence of individual raters, without taking away the ability to predict.

We formulate a set of axioms that we want prediction systems to satisfy. That approach is inspired on social choice theory [15]. Arrow's impossibility Theorem [2] states that the result of a vote must (1) have $X$ over $Y$ if all prefer $X$ over $Y$, (2) let the order of $X$ and $Y$ be independent of $Z$ and (3) there is no dictator. In fact, robustness against manipulation is also a well-studied issue there [15]. Our axioms are fundamentally different, but the idea that certain combinations of axioms do not admit a model is directly taken from social choice theory.

## 3 Axiomatic Requirements

A trust system is robust, when it operates well under attacks. A common way to increase the robustness of the system, is to try to detect attacks. While such detection mechanisms certainly mitigate attacks, they cannot prevent them, by their nature. A detection mechanism detects attacks that have already occurred (at least partially). Ideally, however, we can prevent the attacks from occurring in the first place.

In this paper, we are concerned with attacks that introduce misinformation to the users. However, first, not all attacks induce noise towards the user, but break the system in other ways (e.g. the reputation lag attack [13], where the attacker exploits a time delay before his infamy is spread)[1]. Thus, we only consider attacks by strategic unfair predictions. Second, not all unfair prediction attacks are harmful (e.g. the camouflage attack, where users are honest to gain trust, and betray others when trusted; see [17]). We ignore attacks that do not introduce noise to the user; attacks that do not (aim to) deceive users. Rather, we look at gathering predictions in a robust manner.

[1] We ignore security attacks, such as identity theft or denial of service attacks.

The requirements in this section are informally defined using natural language. Later, we formally define our terminology, and translate the requirements to formal statements. For the sake of precision, we fix the meaning of some terms: A *prediction* is a statement about an *event* before it occurs. An event has an *outcome*, after which the degree of correctness of the prediction is known. *Noise* is the inverse of that degree of correctness. A *rater* is an agent (human or system) that produces predictions. Ratings are *accurate* when they assign the true (subjective) probabilities to outcomes (i.e. outcomes assigned $x\%$ happen $x\%$ of the time), and raters are accurate when they produce accurate ratings. See Sect. 4 for a more formal definition of the terminology.

### 3.1 Robustness Requirements

Consider the strictest formulation of robustness, called absolute robustness (**AR**): **No rater may introduce noise.** Note that **AR** implies that no group of raters may introduce noise either. Intuitively, **AR** seems too strong. If no predictions can introduce any noise, no matter how small or improbable, then how can the rater make any meaningful predictions? In fact, in Sect. 7, we prove this intuition correct; no non-trivial system can be absolutely robust.

The generalisation of absolute robustness is $\theta$-strict robustness (**SR**): **No rater may introduce noise larger than** $\theta$. Note that **SR** implies that no group of raters sized $n$ may introduce noise larger than $\theta \cdot n$. Strict robustness is also a strong axiom, and it is somewhat workable, although it negatively affects other aspects of the system. Particularly, due to the fixed-size tolerance of noise, and the inevitability of noise, any rater can only provide a limited number of predictions. We refer to this property as *effectiveness*, and its axiom is stated below.

Strict robustness can be weakened, e.g. by allowing more noise. Finally, we weaken robustness to $f$-weak robustness (**WR**), where $f$ is a non-decreasing function: **In the first** $n$ **selected predictions, no rater may introduce noise larger than** $f(n)$. Weak robustness is a generalisation of strict robustness (let $f$ be a constant function). Here, good (selected) predictions from the past give the rater some credit. Picking $f(n) = n \cdot \theta$ would encode that the average noise is limited by $\theta$. Whether weak robustness is sufficient is debatable, but we should expect increased effectiveness for some $f$.

We formulate a radically different notion of robustness, based on hypothesis testing. The idea is that one initially assumes perfect accuracy (null hypothesis), and that the null hypothesis may be rejected in favor of *any* alternative hypothesis if the data is unlikely to fit the predictions. The hypothesis testing variant of robustness is (**HR**): **The probability of a sequence of events, given that the predictions are accurate, must not go below** $\alpha$. We show later that **SR** = **HR**, when $\alpha = -2^{\theta}$.

*Remark 1.* Note that we are not subjecting the rater to a single statistical test, but to many. Then, we require that the rater cannot fail any of the statistical tests. This models the notion that we do not know what kind of attack the rater

may be performing (i.e. what the alternative hypothesis is). For every sequence of outcomes there is one statistical test, where $H_0$ is that the rater accurately predicts it, and $H_1$ is that the rater underreports it. For each *individual* statistical test, the probability of falsely rejecting $H_0$ is bounded by $\alpha$. Since we have multiple tests, the probability that at least one test rejects $H_0$ can (and does) exceed $\alpha$.

### 3.2 Auxiliary Requirements

The system that needs to be robust must also have a variety of other properties. The filter should not introduce bias, it must not rely on foreknowledge, and it must not exclude excessively many predictions.

The first requirement is that the system must be implementable – it cannot make decisions based on future events. Specifically, users cannot be prescient (**T**): **Whether a prediction is used should not depend on its outcome, nor on future predictions or outcomes.** There are combinations of requirements that are logically non-contradicting, but that contradict **T**. Rejecting **T** means asserting prescience. Such systems cannot be implemented in a meaningful way, since the purpose of the prediction was to be able to act before the future happened. Note that **T** does not exclude analysing a prediction in the future, it just prohibits users from using such a future analysis in the present.

Another property of a selection mechanism is that it should not be biased. The ideal notion of unbiasedness, called strictly unbiased (**SU**) states: **A prediction from a user about an event is used iff any alternative prediction from that user about that event would be used too.** However, this notion may be too strong, as the mere possibility of an extreme prediction that may introduce an unacceptable amount of noise would imply that all predictions must be blocked. Hence we formulate (weakly) unbiased selection (**WU**): **A prediction from a user about an event is used iff the opposite prediction from that user about that event would be used too.** This notion matches the idea that we can not "prefer" one outcome over the other, and thus that the selection mechanism mistakenly favours one side. However, weak unbiased selection may introduce a bias towards the center, meaning unlikely events may be overestimated.

Finally, the property that forms the typical trade-off with robustness: effectiveness. Effectiveness measures how many predictions can be used over the lifetime of a trust system. We formulate two incarnations of effectiveness. The first is optimistic $k, n$-effectiveness (**OE**): **It is possible to select $k$ predictions for $n$ events.** Optimistic $k, n$-effectiveness can be used to prove hard limits on robustness of trust systems. The second notion of effectiveness is realistic $k, n$-effectiveness (**RE**): **Assuming all raters are accurate, we can expect $k$ predictions for $n$ events.** The realistic $k, n$-effectiveness is used for the positive results.

## 4 Modelling

Raters send predictions to users – be it by broadcasting or upon request. Predictions concern events with an outcome that will eventually be known. Users want to estimate how likely outcomes of events are, and use predictions for this purpose. After the event, users use the outcome to judge the predictors. Good predictors assign high likelihood to actual outcomes, and bad predictors assign lower likelihood.

There is a sequence $P$ of binary events, where the $i^{\text{th}}$ event, denoted $p_i$, either equals 0 or 1. The prediction of rater $a$ about $p_i$ is $r_i^a$, which (for honest raters) represents his estimate of $p(p_i{=}1)$ – and $\overline{r_i^a} = p(p_i{=}0) = 1 - p(p_i{=}1)$. The sequence of all predictions of rater $a$ is $R^a$, with $R_i^a$ his prediction about the $i^{\text{th}}$ event. For a set of raters $A$, we can write $R^A$ to mean $\{R^a | a \in A\}$ Together $A$, $P$ and $R^A$ form a trust system.

The user has no influence on the values of the predictions or on the outcomes of the events. The only way to achieve the goal of dealing with predictions in a robust manner, is to select the right predictions from the predictions that are given. Note that blocking raters can be accomplished by never selecting that rater's predictions, regardless of the values. Thus, the focus on this paper is on selecting the right predictions. The sequence of predictions that is selected is called the sequence of *filtered* predictions, denoted $\widehat{R}^A$ (where $R^A$ is the set that $\widehat{R}^A$ is selected from).

Our motivating question is what the limitations are to such a filter. The filtered predictions may be biased, can we avoid such a bias? All things considered equal, a looser filter is superior, as it allows the user to consider more information. How many (sufficiently unbiased) predictions can $\widehat{R}$ contain? Finally, the crucial question, can we put a hard limit on how much noise a rater can introduce?

Every prediction has an amount of information [11]. Information is the dual of entropy, and entropy is the expected surprisal of a random variable [5]:

**Definition 1.** *Let $X, Y, Z$ be discrete random variables.*
*The* surprisal *of an outcome $x$ of $X$ is $-\log(P(x))$.*
*The* entropy *of $X$ is $H(X) = \mathbb{E}_X(-\log(P(x)) = \sum_i P(x_i) \cdot -\log(P(x_i))$.*

Once the actual outcome $p_i$ of the event is known, we can analyse the surprisal of the prediction, which is $-\log r_i^a$ or $-\log \overline{r_i^a}$, when $p_i = 1$ or $p_i = 0$, respectively. The surprisal of $r_i^a$ given the outcome $p_i$ is denoted $f'(r_i^a, p_i)$ (to avoid the case distinction for $p_i$). With perfect information (zero entropy), the surprisal is 0, so surprisal measures noise (misinformation).

Therefore, surprisal can be used to measure the quality of a prediction (this is, e.g., the basis of the cross-entropy method [5]). A high quality prediction assigns a higher probability to the actual outcome. But more importantly, a prediction is of low quality when a low probability is assigned to the outcome. Since a high surprisal corresponds to low quality predictions, we use surprisal to measure the noise of a prediction. However, a high degree of noise in the prediction does not necessarily mean that the rater was malicious or even in error. Raters can introduce noise by sheer bad luck.

Other measures could be used than surprisal. There are, however, two advantages being logarithmic: First, the sum of the surprisal of two outcomes of independent events is equal to the surprisal of the outcome of the joint event. The surprisal of a combination of outcomes is the sum of the surprisal of the individual outcomes; formally $\log p(x) + \log p(y) = \log p(x, y)$, for independent $X, Y$. Second, it matches the intuition that the difference between 1 % and 2 % is far more significant than the difference between 50 % and 51 %. However, we also consider another measure for the quality of predictions, which is based on hypothesis testing; a statistical tool, rather than information theoretic.

Before continuing, we define a couple of shorthand notations. Typically, we denote predictions with $r$, but we may use $q$ instead. Furthermore, we allow substitution in a sequence/set, denoted $R^A[r_i^a \setminus q_i^a]$, where $r_i^a$ is replaced by $q_i^a$ in $R^A$. We may want to get the index of $r_i^a$ in the sequence $\widehat{R}$, which we denote as $\rho_{\widehat{R}}(r_i^a)$. Finally, $X \sqsubseteq Y$ if $X$ is a subsequence of $Y$ (same elements and order). These notations are particularly introduced to simplify the notation of the axioms.

## 5 Axioms

We need to have a formal model of the trust system to base the formal version of our axioms on. The idea here follows the standard approach in social choice. We formulate a generic collection of events and predictions, and prove that set of filtered predictions can satisfy a certain combination of axioms. Thus, we can show the impossibility of a combination of desirable properties.

Axiom **AR** – absolute robustness – must encode that "no rater may introduce noise." That axiom can be stated as:

**AR** : $$\forall_{i,a} \sum_{r_i^a \in \widehat{R}^a} f^{\prime}(r_i^a, p_i) = 0.$$

Axiom **SR** – strict robustness – must encode that "no rater may introduce noise larger than $\theta$." That axiom can be stated as:

**SR** : $$\forall_{i,a} \sum_{r_i^a \in \widehat{R}^a} f^{\prime}(r_i^a, p_i) \leq \theta.$$

Axiom **WR** – weak robustness – must encode that "in the first $n$ selected predictions, no rater may introduce noise larger than $f(n)$." That axiom can be stated as:

**WR** : $$\forall_{n,a} \sum_{r_i^a \in \widehat{R}^a \wedge \rho_{\widehat{R}^a}(r_i^a) < n} f^{\prime}(r_i^a, p_i) \leq f(n).$$

Axiom **HR** – hypothesis testing-based robustness – must encode that "the probability of a sequence of events, given that the predictions are accurate, must not go below $\alpha$." That axiom can be stated as:

**HR** : $$\forall_a ( \prod_{r_i^a \in \widehat{R}^a} r_i^a \geq \alpha).$$

The product of the prediction assigned to the actual outcome is what the joint

probability of the outcomes would be if the predictions are accurate. This probability may not go below $\alpha$.

Axiom **T** – non-prescience – must encode that "whether a prediction is used should not depend on its outcome, nor on future predictions or outcomes." The axiom can be stated as:

$$\mathbf{T}: \quad \forall_{i\leq k,a}(r_i^a \in R^a = r_i'^a \in R'^a) \wedge \forall_{i<k}(p_i \in P = p_i' \in P') \implies (r_k^a \in \widehat{R}^a \Leftrightarrow r_k^a \in \widehat{Q}^a),$$

whenever two trust systems are equal up to point $k$, they must allow the same predictions to be selected or blocked. In other words, at time $k$ the selection cannot depend on $p_{k+j}$ or $r_{k+j}^a$, since there exists a system identical up to $k$ steps with $p_{k+j} \neq p'_{k,j}$ and $r_{k+j}^a \neq r_{k+j}'^a$.

Axiom **SU** – strong unbiasedness – must encode that "a prediction from a user about an event is used iff any alternative prediction from that user about that event would be used too." The axiom can be stated as:

$$\mathbf{SU}: \qquad \forall_{i,a}(r_i^a \in \widehat{R}^a \wedge Q^a = R^a[r_i^a \setminus r_i'^a] \implies r_i'^a \in \widehat{Q}^a),$$

every prediction from user $a$ at time $i$ can be replaced by another prediction from $a$ at $i$.

Axiom **WU** – weak unbiasedness – must encode that "a prediction from a user about an event is used iff the opposite prediction from that user about that event would be used too." The axiom can be stated as:

$$\mathbf{WU}: \qquad \forall_{i,a}(r_i^a \in \widehat{R}^a \wedge Q^a = R^a[r_i^a \setminus \overline{r_i^a}] \implies \overline{r_i^a} \in \widehat{Q}^a),$$

every prediction from user $a$ at time $i$ can be replaced by another prediction from $a$ at $i$.

Axiom **OE** – optimistic effectiveness – must encode that "it is possible to select $k$ predictions for $n$ events." The axiom can be stated as:

$$\mathbf{OE}: \qquad \forall_a(\max_{i<n} \rho_{\widehat{R}^a}(r_i^a) \geq k),$$

there highest index of a prediction in $\widehat{R}^a$ with index below $n$ in $R^a$ is at least $k$.

Axiom **RE** – realistic effectiveness – must encode that "assuming all raters are accurate, we can expect $k$ predictions for $n$ events." The axiom can be stated as:

$$\mathbf{RE}: \qquad \forall_{a,\widetilde{R}^a \sqsubseteq R^a}(\max_{i<n} \rho_{\widehat{R}^a[\widetilde{R}^a \setminus \overline{\widetilde{R}^a}]}(r_i^a) \geq k),$$

which is similar to **OE**, except it must also hold if we swap arbitrary values of $r_i^a$ for their negation. With the arbitrary swapping of predictions, **RE** captures the possibility that the actual outcome was $\overline{p_i}$, in which case the surprisal would be $-\log(\overline{r_i^a})$, rather than $-\log(r_i^a)$. Thus, the effectiveness here is attainable for all sequences of outcomes, rather than just one.

## 6 Relative Strength of the Axioms

With the exception of Theorem 1, all the propositions and corollaries in this section are straightforward sanity proofs. Propositions 1, 2, 3 and 4 and Corollaries 1 and 2 merely show that axioms that are supposed to be weaker are indeed weaker. The relative strength of the axioms is depicted in Fig. 1.

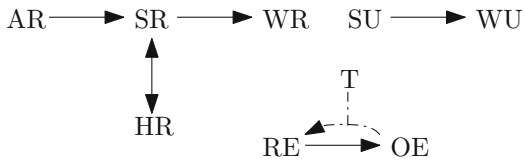


**Fig. 1.** Relations between axioms. Arrows point from strong to weak.

Theorem 1 is the only deep result in this section, as it shows the equivalence between $\mathbf{SR}(\theta)$ and $\mathbf{HR}(\alpha)$, for $\alpha = \frac{1}{2^\theta}$. Thus Theorem 1 shows that an information-theoretic perspective coincides with a view based in statistical methods; specifically hypothesis testing.

The first proposition shows that a lower fixed robustness threshold is a stronger requirement:

**Proposition 1.** *If* $\theta \leq \theta'$, *then* $\mathbf{SR}(\theta) \implies \mathbf{SR}(\theta')$.

*Proof.* By transitivity: $\forall_{i,a} \sum_{r_i^a \in \widehat{R}^a} f^{\mathbf{f}}(r_i^a, p_i) \leq \theta \leq \theta'$.

Proposition 1 shows that strict robustness is a weaker requirement than absolute robustness:

**Corollary 1.** *For all* $\theta$, $\mathbf{AR} \implies \mathbf{SR}(\theta)$.

The second proposition shows that a consistently lower robustness threshold is a stronger requirement:

**Proposition 2.** *If, for all* $n$, $f(n) \leq f'(n)$, *then* $\mathbf{WR}(f) \implies \mathbf{WR}(f')$.

*Proof.* By transitivity: $\forall_{i,a} \sum_{r_i^a \in \widehat{R}^a} f^{\mathbf{f}}(r_i^a, p_i) \leq f(n) \leq f'(n)$.

Proposition 2 shows that weak robustness is a weaker requirement than strict robustness:

**Corollary 2.** *If* $f(1) \geq \theta$, *then* $\mathbf{SR}(\theta) \implies \mathbf{WR}(f)$.

The third proposition shows that no bias towards any prediction is a stronger requirement than no bias w.r.t. the opposite prediction:

**Proposition 3.** $\mathbf{SU} \implies \mathbf{WU}$

*Proof.* The term $\overline{r_i^a}$ in **WU** is an instance of $r_i'^a$ in **SU**, and **SU** dictates that substitution can be done for all $r_i'^a$.

The fourth proposition shows that realistic effectiveness is a stronger requirement than optimistic effectiveness:

**Proposition 4.** $\mathbf{RE} \implies \mathbf{OE}$

*Proof.* In **RE**, we can find **OE** by letting $\widetilde{R}^a = \emptyset$.

Finally, this section's main theorem, which shows the deep link between information-theoretic robustness and hypothesis testing robustness:

**Theorem 1.** *If* $\alpha = \frac{1}{2^\theta}$, *then* $\mathbf{SR}(\theta) \leftrightarrow \mathbf{HR}(\alpha)$.

*Proof.* Note that $\prod_{r_i^a \in \widehat{R}^a} r_i^a \geq \alpha$ iff $\log(\prod_{r_i^a \in \widehat{R}^a} r_i^a) \geq \log(\alpha)$. Distributing the log over the product and negating, $-\sum_{r_i^a \in \widehat{R}^a} \log(r_i^a) = -\log(\alpha)$. This is $\mathbf{SR}(\theta)$ with $\theta = \log(1/\alpha)$.

## 7 Impossibility Results

Here, we study the relationship between the axioms. Specifically, we investigate whether certain combinations of axioms admit a non-trivial set of filtered ratings. Moreover, where applicable, we investigate what the size of the set of filtered ratings can be. Any statement made in this section is a general truth about all rating systems. The results are summarized in Table 1.

The first is the effective impossibility of a system that has absolute robustness. The only ways in which a system can be absolutely robust, is if it either never uses predictions or if it only uses predictions that predict 100 % probability for the correct outcomes. The former implies an effectiveness of 0 (i.e. it is *ineffective*); the latter breaks non-prescience. An absolutely robust trust system without prescience is ineffective:

**Theorem 2.** $\mathbf{AR} + \mathbf{T} + \mathbf{OE}(k, n) \implies k = 0$

*Proof.* Let $\widehat{R}^a$ be a subset of $R^a$ such that it satisfies **AR**. Since noise is a positive quantity, $\forall_{r_i^a \in \widehat{R}^a} f'(r_i^a, p_i) = 0$. Thus $r_i^a = 1$ iff $p_i = 1$. If $\widehat{R}$ is non-empty, then we can take such $a, i$. Due to **T**, when $p_i = 0$, $\widehat{R}$ remains the same up to $i$. However, if $p_i = 0$, then $r_i^a = 1$ implies $f'(r_i^a, p_i) = \infty > 0$. By **T**, if $\widehat{R}$ is non-empty, then there exists a system that violates **AR** or **T**. Hence, $\widehat{R} = \emptyset$ and $k = 0$.

This theorem (and the following) can be stated as an impossibility theorem: **There is no non-prescient, effective, absolutely robust trust system.**

Moreover, even if we drop non-prescience (thus using predictions given foreknowledge), the system would not even be weakly unbiased unless all predictions are ignored. In other words, if we select 100 % correct predictions, we would lose (weak) unbiasedness. A weakly unbiased absolutely robust trust system is ineffective:

**Proposition 5.** $\mathbf{AR} + \mathbf{WU} + \mathbf{OE}(k, n) \implies k = 0$

*Proof.* Similar to to Theorem 2, except rather than swapping $p_i$, we swap $r_i^A$.

**There is no unbiased, effective, absolutely robust trust system.**

Weakening the robustness requirement to strict robustness, we finally obtain a bit of robustness. A non-prescient rating system with strict robustness can allow at most $\theta$ ratings to be selected from users:

**Theorem 3.** $\mathbf{SR}(\theta) + \mathbf{T} + \mathbf{OE}(k, n) \implies k \leq \theta$

*Proof.* Let $\widehat{R}^a$ be a sequence of $r_i^a$. Due to axiom **T**, the choice of $r_i^a$ is independent of $p_i$. Thus, if $r_i^a \neq 1/2$, then the model must hold with noise $f^{\prime}(r_i^a, p_i)$ and $f^{\prime}(r_i^a, \overline{p_i})$. Without loss of generality, we can therefore assume $r_i^a \leq \overline{r_i^a}$. Now, via $\mathbf{SR}(ic)$, $\theta \geq \sum_{r_i^a \in \widehat{R}^a} f^{\prime}(r_i^a, p_i) \geq \sum_{r_i^a \in \widehat{R}^a} f^{\prime}(1/2, p_i) = k$

**There is no non-prescient, unboundedly effective, strictly robust trust system.**

An interesting academic question is whether the fixed bound on effectiveness can be lifted when we are aware of the future. It turns out that if we replace non-prescience with weak unbiasedness, that the bound is widened, but still fixed:

**Theorem 4.** $\mathbf{SR}(\theta) + \mathbf{WU} + \mathbf{OE}(k, n) \implies k \leq 2^{\theta}$

*Proof.* As **T** is not an axiom, we can select $r_i^a$ knowing $p_i$. However, due to **WU**, $f^{\prime}(\widehat{R}^a_{\leq k}, p) + \overline{r_i^a}$ must be at most $\theta$. Let $c_k = \theta - f^{\prime}(\widehat{R}^a_{\leq k}, p)$. Then we obtain the recursive equation $c_k + \log(1 - \frac{1}{2^{c_k}}) = c_{k-1}$. Via $1 - \frac{1}{2^{c_k}} = 2^{c_{k-1} - c_k}$, and $2^{c_k} - 1 = 2^{c_{k-1}}$ that becomes $c_k = \log(2^{c_{k-1}} + 1)$. Basic arithmetics show that $c_k = \log(k)$.

**There is no unbiased, unboundedly effective, strictly robust trust system.**

When tightening the requirement on unbiasedness to strong unbiasedness, we lose effectiveness completely. Even without non-prescience. Thus, strict robustness and strong unbiased cannot be meaningfully combined.

**Theorem 5.** $\mathbf{SR}(\theta) + \mathbf{SU} + \mathbf{OE}(k, n) \implies k = 0$

*Proof.* Let $r_i^a$ in $\widehat{R}^a$, then the theorem must also hold for $q_i^a$. However, if we let $q_i^a < -\log(\theta)$, and $p_i = 1$, then the strict robustness is broken. Hence, there cannot be any $r_i^a \in \widehat{R}^a$, and $k = 0$.

**There is no strongly unbiased, effective, strictly robust trust system.**

Again, we weaken the robustness requirement. When we keep strong unbiasedness, we again lose effectiveness. Thus, not a single notion of robustness can combine meaningfully with strong unbiasedness.

**Theorem 6.** $\mathbf{WR}(\theta) + \mathbf{SU} + \mathbf{OE}(k, n) \implies k = 0$

*Proof.* Reuse the proof of Theorem 5, replacing $\theta$ with $f(1)$.

**There is no strongly unbiased, effective, weakly robust trust system.**

Finally, we consider a weakly robust, non-prescient system. Here, the limitation on the effectiveness is the weakest (assuming $\theta = f(1)$):

**Theorem 7.** $\mathbf{WR}(f) + \mathbf{T} + \mathbf{OE}(k, n) \implies k \leq f(n)$

*Proof.* Reuse the proof of Theorem 3, replacing $\theta$ with $f(n)$.

**There is no non-prescient, unlimited effective, weakly robust trust system.**

## 8 Robust Prediction Systems

We have shown the negative impact of robustness on other desirable requirements on a rating system. Perhaps robustness is simply a problematic notion in itself. In the proofs, we have shown that models cannot exist in certain combinations, not that models do exist in the negation. In this section, we show that there do exist reasonable models that strike a balance between robustness, fairness and effectiveness.

It does not suffice to prove the converse of the impossibility theorems, as that would simply prove that there exists a set of filtered ratings of a certain size. However, the setting in which that size is reached may be a pathological case. We want to show that filters can be *expected* to achieve a certain size. Hence, we are using axiom **RE**, rather than **OE**. Realistic effectiveness is an assertion about raters whose ratings correspond to the true probabilities. We want to show that these honest raters are expected to have a certain number of ratings selected by the filter.

First we introduce an auxiliary lemma, that shows that under **T**, **RE** = **OE**:

**Lemma 1.** $\mathbf{T} + \mathbf{RE}(k, n) \Leftrightarrow \mathbf{T} + \mathbf{OE}(k, n)$

*Proof.* For given $\widehat{R}^a$, if $r_i^a$ is in $\widehat{R}^a$, then, via **T**, $\overline{r_i^a} \in \widehat{Q}^a$. We can take $\widetilde{R}^a$, and swap all $r_i^a$ as above. Then we can make **OE** match any individual instance of **RE** for $\widetilde{R}^a$. Thus, for $\mathbf{OE}(k, n)$ and **T** to hold, $\mathbf{RE}(k, n)$ can be deduced to hold too. Together with Proposition 4, that proves the lemma.

The following theorem concerns strict robustness. A non-prescient, weakly unbiased, strictly robust filter can be expected to have over $\theta$ ratings selected over a lifetime:

**Theorem 8.** *There is a model that satisfies* $\mathbf{T} + \mathbf{WU} + \mathbf{SR}(\theta) + \mathbf{RE}(\theta - 1, n)$*, for sufficiently large* $n$*.*

*Proof.* Via Lemma 1, it suffices to prove for $\mathbf{T} + \mathbf{WU} + \mathbf{SR}(\theta) + \mathbf{OE}(\theta - 1, n)$. If we only select those ratings that are within distance $\epsilon$ from $1/2$, then the noise is at most $k + k \cdot \epsilon$. Letting $k = \theta - 1$, the noise is at most $\theta - 1 + (\theta - 1) \cdot \epsilon$, which is under $\theta$ for sufficiently small $\epsilon$. It is straightforward to verify that this scheme does not violate **WU**.

For the next theorem, we look at an interesting subclass of weak robustness. We consider only those functions where $f(i) - f(i-1)$ is constant; specifically, 1. Thus, every prediction, the rater gets an additional bit credit. If the rater randomly provides ratings, the expected loss equals the gain and a bad rater is expected to make a nett loss. Specifically, the expected change in nett score is $f'(r_i^a, p_1) - 1$, which can be negative, 0 or positive.

**Theorem 9.** *Raters whose ratings do not correlate with the events, or correlate negatively, have a finite effectiveness. Raters whose ratings correlate positively with the event have a non-zero probability of infinite effectiveness.*

*Proof.* This is a simple application of a rule in random walks [7]. The probability of ruin – losing all credit – is 1 for random walks with $\mathbb{E}(step) \leq 0$, and the probability of ruin is strictly below 1 for random walks with $\mathbb{E}(step) > 1$.

Theorem 9 is a superficially surprising result. We have a hard guarantee that below average predictors are eventually unable to get their ratings selected. We cannot guarantee that a high quality predictor is not shunned too. If a high quality predictor is unlucky, he can still have a random walk ending in ruin. Note that for simplicity, we took the cutoff at $1/2$, we could have chosen arbitrary values, or even a dynamic version. In all these cases, random walks without expected gain eventually run into ruin.

## 9 Conclusion

We have presented a simple formal model for prediction systems. That formal model focusses on the actual predictions, the outcomes and which predictions are used, and ignores the non-essential aspects of a prediction system.

We have outlined desirable properties for such a prediction system to have. Specifically, we have three notions of robustness – how much noise an attacker (or any rater) can introduce – in various strength (absolute, strict, weak). All three notions are formulated in information theory, but strict robustness can be stated in classical statistical term too. Two notions deal with bias, the strong version disallows any form of bias, whereas the weak version allows bias towards the center. Two more notions deal with effectiveness – how often the user can use the ratings. One version (weak) overestimates the effectiveness, to strengthen the impossibility results. Finally, one notion deals with the fact that users should not be able to foreknow the future.

All these notions have been translated into axioms in the language of the formal model for prediction systems. We show that the axioms do indeed satisfy the desired strength relations.

Based on the axioms, we present a collection of impossibility results. The absolute notion of robustness cannot have any effectiveness whatsoever. The strict notion of robustness can have a bounded effectiveness, meaning that the system cannot keep providing useful predictions indefinitely. For the weak notion of robustness the effectiveness remains hampered.

Finally, we show that a strict robust system can exist, and while its life-span is limited, reaching the theoretically maximal effectiveness is feasible. More importantly, we show that if we weaken robustness, an interesting property regarding effectiveness arises. Selecting the right function ($f(n) = n + \theta$), we get that better-than-random raters could have infinite effectiveness, whereas random-or-worse raters have finite effectiveness. In other words, the quality of the ratings determines whether the effectiveness is bounded.

Together the results in this paper sketch the idea that fairly strong notions of robustness are feasible, but come at a high cost. An interesting research direction would be to fine-tune all the desirable properties into an actual system – rather than a theoretically induced model.

## References


1. Arrow, K.J., Forsythe, R., Gorham, M., Hahn, R., Hanson, R., Ledyard, J.O., Levmore, S., Litan, R., Milgrom, P., Nelson, F.D., et al.: The promise of prediction markets. Science **320**(5878), 877 (2008). New York, Washington
2. Arrow, K.J.: Social Choice and Individual Values. Cowles Foundation Monographs Series. Yale University Press, New Haven (1963)
3. Berg, J.E., Nelson, F.D., Rietz, T.A.: Prediction market accuracy in the long run. Int. J. Forecast. **24**(2), 285–300 (2008)
4. Buckley, P., OBrien, F.: The effect of malicious manipulations on prediction market accuracy. Inf. Syst. Front. 1–13 (2015). http://link.springer.com/article/10.1007/s10796-015-9617-7
5. Cover, T.M., Thomas, J.A.: Entropy, relative entropy and mutual information. In: Elements of Information Theory, pp. 12–49 (1991)
6. Deck, C., Porter, D.: Prediction markets in the laboratory. J. Econ. Surv. **27**(3), 589–603 (2013)
7. Feller, W.: An Introduction to Probability Theory and Its Applications, vol. I. Wiley, London, New York, Sydney, Toronto (1968)
8. Green, K.C., Armstrong, J.S., Graefe, A.: Methods to elicit forecasts from groups: Delphi and predictionmarkets compared (2007)
9. Hanson, R., Oprea, R., Porter, D.: Information aggregation and manipulation in an experimental market. J. Econ. Behav. Organ. **60**(4), 449–459 (2006)
10. Hoffman, K., Zage, D., Nita-Rotaru, C.: A survey of attack and defense techniques for reputation systems. ACM Comput. Surv. **42**(1), 1 (2009)
11. Jianshu, W.E.N.G., Chunyan, M.I.A.O., Angela, G.O.H.: An entropy-based approach to protecting rating systems from unfair testimonies. IEICE Trans. Inf. Syst. **89**(9), 2502–2511 (2006)
12. Jøsang, A., Ismail, R., Boyd, C.: A survey of trust and reputation systems for online service provision. Decis. Support Syst. **43**(2), 618–644 (2007)
13. Kerr, R., Cohen, R.: Smart cheaters do prosper: defeating trust and reputation systems. In: Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems, vol. 2, pp. 993–1000. International Foundation for Autonomous Agents and Multiagent Systems (2009)
14. Snowberg, E., Wolfers, J., Zitzewitz, E.: Partisan impacts on the economy: evidence from prediction markets andclose elections. Technical report, National Bureau of Economic Research (2006)
15. Taylor, A.D.: Social choice and the mathematics of manipulation. Cambridge University Press, Cambridge (2005)
16. Wang, D., Muller, T., Irissappane, A.A., Zhang, J., Liu, Y.: Using information theory to improve the robustness of trust systems. In: Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), pp. 791–799 (2015)
17. Wang, D., Muller, T., Zhang, J., Liu, Y.: Is it harmful when advisors only pretend to be honest? In: Proceedings of the 30th AAAI Conference on Artificial Intelligence (2016)