In response to the reviews received from our IJCAI 2015 submission, we have rewritten many parts of the paper to make things more clear. In a nutshell:

1. We have done more experiments. We have tested the method we propose on a large number of games. There is no space in the paper to include all the results, but we have created an additional document with the details on the different games. It is available for anonymous download at: http://www.megafileupload.com/9blI/supplement.pdf.

2. We have added a new class of opponents, an opponent that switches strategy at some point in the repeated play.

3. We have cleaned up the text in many places, trying to make it more clear.

——— Review from Reviewer 1 ————
Originality of ideas : 7
Correctness : 7
Significance of results : 3
Clarity and quality of presentation : 3
Confidence : 7
Overall recommendation : 7

Comments to the author(s):
The paper proposes a new solution concept for balancing between exploiting a prediction and guarding against best-responding or worst-case opponents in general-sum games. Empirically they demonstrate that the new approach, called Restricted Stackelberg Response with Safety, improves performance in a small 2-player general-sum game against an omniscient best-responding opponent and against a worst-case opponent, over using prior approaches of Restricted Nash Response and Safe Policy Selection.

This is a novel and very interesting idea. But I am not left with a clear takeaway for when this would be useful or preferable to other approaches. What domains outside of the small game considered would you want to use this in? How would it do against unknown opponents other than a pure omniscient best response or absolute worst case opponent? I am not left with a bigger picture perspective other than that it helped against these two classes in one particular game.

Thanks for noticing the novelty of the ideas we proposed. We have added more explanations on potential domains of application. We have also obtained results on a large set of games that are consistent with the game reported in the paper. We cannot include them for lack of space. We have also added another type of opponent, an opponent that switches strategies after 50 rounds.

I think the idea is interesting and has a lot of promise, but there are a lot of things that I think must be addressed and clarified before publication. I could be persuaded to change my mind however if the rebuttal addresses my main criticisms adequately.

Abstract says 'We describe an algorithm which appropriately selects parameter values and show experimental results in a general-sum game against a variety of opponents.’ But the results are only against two opponents ‘a best-responding opponent and a worst-case opponent. I wouldn’t call this a variety of opponents.

Thanks for the suggestion. We have corrected the abstract and added one more type of opponent.
They should clarify that the work is specifically for two-player general sum games. The analysis does not seem directly applicable to more than two agents, though it could perhaps be extended.

Done.

It appears that the results would extend to extensive-form games, with perfect and imperfect information, but there is no discussion of this extension (though extensive-form imperfect-information games was the setting that RNR was developed for and implemented on, as well as some of the other approaches cited in the introduction.)

We have focused on normal form games for our initial research, but we agree that extensive-form imperfect-information games are a logical next step.

I think it should be clarified that the prediction is a single point estimate, and not a distribution over strategies. Would the approaches be applicable to the more general Bayesian setting where the posterior opponent model is a full distribution? If so, how would the analysis change? These two approaches are equivalent in the strategic-form case of stationary strategies with Dirichlet prior, but not in general, and if the Dirichlet prior assumption is crucial to this analysis it should receive more emphasis in the description of the model.

The Dirichlet prior is only a function of the prediction algorithm we’re using – it’s not an intrinsic property of RSRS. We have clarified that the prediction is a single point estimate – a distribution over moves, not a distribution over mixed strategies. When we extend RSRS to extensive form games we will take a closer look at how predictions of opponent behavior should be described.

In the introduction: Playing a maximin strategy avoids exploitation, but at the cost of performance against the prediction.

Exploitation is not well defined in general-sum and multiplayer games. For two-player zero-sum games, the exploitability is the difference between the value of the game and performance against a nemesis. But there is no notion of a value in general-sum games. I don’t see any way in which playing a maximin strategy avoids exploitation. In fact, in a sense you are being precisely exploited if you are forced to play an ultra-conservative maximin strategy to protect against the absolute worst-case scenario. I have never seen maximin proposed as a serious solution concept to be implemented by agents. I would be interested in seeing citations for prior work that uses it, and would also like to see the formal definition of ‘exploitability’ for general-sum games that is proposed.

We were using ‘exploit’ in an informal sense, and we regret implying that there was a formal definition. We have clarified our meaning in the paper. That said, while it is not desirable to play against an opponent for which it is necessary to adopt the maximin strategy, we would expect a rational agent to do so, rather than accept a worse outcome.

Playing a Nash equilibrium is equivalent to discarding the prediction in favor of assuming the opponent will play the Nash equilibrium. I don’t see how the Nash equilibrium discards the prediction any more than the maximin strategy does. That said, it is true that NE/maximin strategies don’t account for exploiting a possible prediction. But I would try to elaborate and explain this better.

We have changed the phrasing to be clear that playing either maximin or a Nash equilibrium ignores the value of the prediction.
I’d like to see justification for why we are concerned specifically for on the possibility that the opponent is a best responder or worst case. How might we do against approximate best-responders, or opponents who are playing far away from the prediction but not best responding?

We focused on those two options because we felt they were the most significant threats. Our weight learning algorithm was developed to handle approximate best responders. Opponents which are playing far from the prediction can only be handled by improving the predictive algorithm. For example, in a 10 move game an opponent which plays the digits of pi in sequence will be impossible to predict if the existence of such an opponent had not been anticipated. Fictitious play will (after sufficient observations) predict a uniform distribution over digits, even though the opponent is playing a pure strategy each round. We don’t feel this problem needs to be addressed because there is no incentive for an opponent to play in such a way – we feel that, while it is regrettable that our agent would not achieve the best possible payoff, the performance of the best response to a uniform distribution would be the best one could reasonably hope to achieve in such a situation.

Make it clear that the concept of a game value is not well defined for non-zero-sum games.

We have changed the wording to make it clear that the value we are talking about is not the value of the game.

Stackelberg equilibrium not introduced in background section even though it is central concept to paper.

We have added a brief discussion and reference to the Related Work section.

It would be helpful for a sentence providing some intuition for the demonstration game beyond the rules, since this is the only game considered for motivation and the experiments.
There is a variant called Rock-Paper-Scissors-Lizard-Spock from the popular tv show Big Bang Theory that has totally different payoffs from the Rock-Spock-Lizard-Paper-Scissors game described. This should be mentioned. Should also say that the game is not zero-sum and symmetric, and should state what the unique equilibrium is (and justification for why it is unique).

We have tested the method on multiple other games. We did not report the results for lack of space. They are consistent with the results in the game we describe in detail.

“In general-sum games, increasing the weight can reduce performance against the prediction.” Please provide some explanation. In zero-sum games does performance increase monotonically with the weight, and why?

We have not observed an example of a zero-sum game where performance against the prediction was reduced when the weight was increased, and we suspect that it is not possible, but we don’t have a formal proof.

“Nash equilibria are not necessarily unique in general-sum games.” – They aren’t necessarily unique in zero-sum games either.

We agree. However, in a zero-sum game, Nash equilibria will all be members of a single connected component of Nash equilibria - they will all be interchangeable - this is not true in a general-sum game.

The authors say that the RNR-game formulation is general-sum: it isn’t clear to me that it isn’t zero sum, and I’d like to see this discussed. They also state that RNR applies to zero-sum games, but it isn’t clear why it wouldn’t also apply to general-sum games (in fact, the experiments seem to apply it to a general-sum game).
The modified game created to calculate a RNR only changes the payoffs for one of the players. It is possible for the modified game to be zero-sum (depending on the prediction) but not inevitable. RNR can easily be applied to general-sum games – the only difficulty lies in selecting a Nash equilibrium when there are multiple equilibria available.

The definition of “distinctness” of Nash equilibrium is not clear at all to me: “Two equilibria are considered distinct if each player strictly prefers to play their equilibrium strategy in an equilibrium. They should define this concept formally.

We have provided a formal definition in the paper. Two equilibria s and s’ are considered distinct if 
\[ U_1(s_1, s_2) > U_1(s_1', s_2'), U_1(s_1', s_2') > U_1(s_1, s_2'), U_2(s_1, s_2) > U_2(s_1, s_2'), \text{ and } U_2(s_1, s_2') > U_2(s_1', s_2). \]

Another way to phrase it is to say two equilibria are considered distinct if they are members of separate connected Nash components.

I dont follow the argument the general-sum RNR formulation has a unique equilibrium. I also dont see how it would be true. E.g., what about a trivial game where payoffs are always zero. Then any strategy would be an RNR for any prediction/weight, and there would be multiple equilibria for any formulation.

When we speak of multiple equilibria we mean equilibria in separate connected components. If both players select different equilibria from the same connected component they will receive the same payoff as they would have if they had selected the same equilibrium. In the example you give, both players will receive a payoff of 0 regardless of whether they select the same equilibrium or not, and as a result it doesn’t matter which equilibrium they select. Similarly, when calculating a RNR if there is only one connected component of equilibria it doesn’t matter which equilibrium you select from that connected component.

“[SPS] does this by gradually increasing the r value, but reducing it whenever the agent receives a payoff less than the value of the game ..” No, it bases the increase/reduction on the expected payoff, using randomization of our action and chance, not over the actual payoff it received.

We have corrected our misstatement.

It is not surprising to me that the performance was generally better against prediction than against best response than against worst-case. But I did find it interesting that there was a reasonably-sized interval in the w space for which RNR vs. best response actually outperformed RNR vs. the predicted strategy.

This was quite surprising for us as well. We believe it can occur when the benefits of being a Stackelberg leader outweigh the benefits of performing well against the prediction.

Would be helpful if ”Stackelberg leader” were defined, and the benefits of being a Stackelberg leader were discussed.

We have included a brief discussion of Stackelberg leaders.

Would be interesting to see some more discussion about the discontinuities in the plots. Will these always be continuous for all the opponent types/are there settings where they will be continuous?

Changes in r will cause continuous changes in the strategy. Changes in w cause discontinuous changes. Our algorithm finds a Stackelberg solution, so it always selects a strategy such that the opponent’s best response is a specific move. The discontinuities occur where the changing w values result in a different move by the opponent.
100 rounds doesn’t seem like very long for a match. Is there a reason $T = 100$ was selected over 1,000 or 10,000?

We selected 100 rounds because it provided ample time for the (simple) predictor to develop a (possibly inaccurate) prediction of the opponent.

Comments after responses:
I think that most of my comments were pretty minor, and several were addressed adequately in the responses. My biggest issue was with correctness of the result in Section 3. My confusion stemmed from several issues that I outline below. I think the authors should make the following modifications:

1) They have a typo in equation 2, and the first $s_1'$ in the second line should be $s_1$ without a prime.

Thanks for noticing it. We fixed it.

2) They explained distinct Nash equilibrium poorly. They said "Assume $G'$ has two distinct Nash equilibria $s$ and $s'$. Two equilibria are considered distinct if each player strictly prefers to play their equilibrium strategy in equilibrium. If the preference is weak it doesn’t present the same dilemma for RNR because the players can play either strategy and achieve the same payoff.

They should have said something like this instead: "Assume $G'$ has two distinct Nash equilibria $s$ and $s'$. A Nash equilibrium is distinct if each player strictly prefers to play their equilibrium strategy in their equilibrium [should provide formal definition]. If the preference is weak it doesn’t present the same dilemma for RNR because the players can play either strategy and achieve the same payoff."

Thanks for the suggested wording, which we adopted.

They are defining distinctness for a single Nash equilibrium, but the wording makes it sound like they are defining a relation between two different Nash equilibrium (by wording it "Two equilibria are considered distinct if...")

Distinctness is a relation between two different equilibria. We regret the poor wording in the previous version and have improved it.

The proof is correct with the modifications I have mentioned, and I think that result is very interesting actually. They should definitely make the result in Sect 3 a theorem, and should define “distinct Nash equilibrium” formally as described above.

I’d still like them to account for my other comments too, but many of them can be addressed easily just by wording things differently.

There are also a few related papers on safely exploiting predictions of opponents that I think should be mentioned – "Safe opponent exploitation" by Ganzfried/Sandholm and "A general criterion and an algorithmic framework for learning in multi-agent systems.” by Powers/Shoham/Vu.

We added the citations suggested.
Summary: The paper proposes a new solution concept for balancing between exploiting a prediction and guarding against best-responding or worst-case opponents in general-sum games. Empirically they demonstrate that the new approach, called Restricted Stackelberg Response with Safety, improves performance in a small 2-player general-sum game against an omniscient best-responding opponent and against a worst-case opponent, over using prior approaches of Restricted Nash Response and Safe Policy Selection.

This is a novel and very interesting idea. But I am not left with a clear takeaway for when this would be useful or preferable to other approaches. What domains outside of the small game considered would you want to use this in? How would it do against unknown opponents other than a pure omniscient best response or absolute worst case opponent? I am not left with a bigger picture perspective other than that it helped against these two classes in one particular game.

We added other classical games to show that the method we propose is applicable and works well in more than one game. This research is intended to be useful in any game for which the calculations are not computationally intractable. There are many different approaches to predicting opponent behavior, but if you just best-respond to the prediction your algorithm will always be vulnerable to some other algorithm. You can extend any prediction algorithm with RSRS to make it resistant to best-responding opponents and worst-case opponents - two of the most dangerous opponents for an agent to encounter.

I think the idea is interesting and has a lot of promise. I think there are a lot of things that must be addressed. Specifically, the modifications I described for Section 3, and the other wording and explanation comments I have made.

Review from Reviewer 2:

Originality of ideas: 6
Correctness: 9
Significance of results: 5
Clarity and quality of presentation: 8
Confidence: 7
Overall recommendation: 5

Comments to the author(s):

In repeated games having a prediction of the opponent is useful to select next action. However, that prediction may be incorrect. The authors propose the Restricted Stackelberg Response with Safety (RSRS) as a method that can use a prediction method and at the same time it can makes some guarantees for safe policies for not being exploited.

Their proposal is a combination of previous methods such as Restricted Nash Response (RNR) and Safe policy selection (SPS). RNR uses a weight to the prediction, with the maximum weight it uses best response to the prediction on the other extreme it plays Nash equilibrium. RSRS takes this idea and changes one aspect, instead of playing Nash equilibrium assume the opponent will best respond.

SPS proposes strategies that are r-safe since in worst case they will achieve a payoff of within r of the safety value. RSRS reuses this idea to guard against worst case opponents. Thus, RSRS has two parameters: 1) r that controls the risk willing to accept against a worst case outcome and 2) w, a weight on the prediction.

The authors address that setting the w parameter is difficult and propose a learning method for solving this problem. Experiments are performed on a modified version of Rock/Scissors/Papers where comparisons are made with SPS and RNR.
The paper is well written and provides some theoretical results but it lacks sufficient experimental evidence. These are performed only in one game under certain opponents. Experimental section is compact and fails to provide some conclusions about RSRS and its results.

We did more experiments on many additional games and with one additional opponent type.

Specific questions and comments:
The authors use fictitious play as predictor, is the method still useful with another predictor?

Our method is intended for use with any predictor. We used fictitious play as a simple example.

Related work does not mention any works in this area for example:
- When speed matters in learning against adversarial opponents
- A framework for learning and planning against switching strategies in repeated games

Thanks for the suggestions. We added the citations.

Experimental results
- Test the method in non-zero-sum games (e.g., prisoner’s dilemma)
- In Figure 4 what will be the score of RSRS with different parameters?
- What happens when with learning opponents? Is the approach still useful?
- Test against more opponents.
- It is not clear how the learning weight algorithm is evaluated in Section 7.

We tested our method on a non-zero sum game, a variation of Rock/Spock/Paper/Lizard/Scissors.
In figure 4, the RSRS agent is selecting its parameters according to the weight-learning algorithm and safe policy selection. If it adopted fixed parameters instead, it would perform worse, because it wouldn’t be using appropriate parameter values for the opponents it is facing.
If the opponent is learning the method is still applicable, because the agent can still try to predict the opponent.
We did not play against a learning opponent, but we are planning on studying self-play.
We have tested against an opponent that switches strategy.
The learning weight algorithm is tested indirectly by testing the performance of an agent using RSRS against multiple opponents.
- Summary: The authors propose a new approach RSRS for generating strategies in normal form games when a predictor is available but takes into account that predictions may be incorrect and assumes that the opponent may best respond. Moreover, they add a safe mechanism to be used against worst-case opponents with guarantees of certain payoff.

RSRS builds upon previous work (RNR and SPS) and joins them into one algorithm.
The paper provides proof of when this algorithm may work best and provides experimental result in a modified Rock-Scissors-Paper game.
The method is novel and theoretical results are given, the main limitation is that more experimental results would be desired.

We have done more experiments, testing the approach on a large number of games. We have also added another opponent type.

——— Review from Reviewer 3 ————
Originality of ideas: 7
Correctness: 6
Significance of results: 6
The paper develops a new solution concept for safely exploiting predictions in games using concepts from Restricted Nash Equilibria and Safe Policy Selection called Restricted Stackelberg Response with Safety (RSRS). They test this solution method on a general-sum version of RPSLS. Although the authors do not explicitly state why they needed to use a modified version of this game, they do this because RNR requires transforming a zero-sum game, like RPS, into a general-sum game. Then they could use the weight parameter they found. They also find the risk parameter from SPS which they explain as a method of exploiting a prediction of opponent behavior in RPS although they do not explain if this was meant to be used with the zero-sum version of RPS or the general-sum modified version.

We meant to use it with the general-sum modified version of RPS.

After providing some empirical results for finding these parameters they provide a proof that reducing the weight parameter improves performance against a best-responding opponent. They also provide a method for calculating the probability that an opponent will use a best-response.

The last part of the paper performs a tournament of agents in fictitious play over 100 games with mixed results and graphs that are incredibly difficult to read.

RELEVANCE:
This work is relevant.

ORIGINALITY:
Much of the contribution of this paper is a combination of known strategies but done in a new way and the proof provides support for the novelty. There have been other similar works in the multi-agent learning literature that use various concepts of targeted optimality and safety, but the specific combination use here appears to be novel.

Thanks for recognizing the novelty of our proposed method.

The paper should also reference the 2002 JAIR article by Tennenholz on Competitive Safety Analysis: Robust Decision-Making in Multi-Agent Systems

Thanks for the suggestion, we have added the citation.

SIGNIFICANCE:
The basic idea of balancing safety and exploiting imperfect predictions about the play of an opponent is a good one, and this problem should probably receive more attention in the game theory community. However, the weak evaluation of the work limits the significance.

Thank you for your positive comments about the work. We have expanded significantly the evaluation part.

TECHNICAL QUALITY:
The basic idea of the approach is fairly straightforward, and draws on several existing ideas. The technical results do not have any immediate errors, though it is not clear what the point of some of this analysis is (lemma 1 in particular seems obvious and not worth the space in a short paper). There is some inconsistency in how the approach is motivated, and whether this is intended as a learning method or a method for playing
in 1-shot games. The evaluation and section 6 suggest that this is really a type of learning method for repeated play, but it is not really evaluated using the criteria typical of these approaches (e.g., convergence behavior, play against various other learning strategies).

*The method is not a learning method but it is a meta-method to be used by an agent on top of a prediction method to decide the action to take next. We show its use in repeated play to quantify the impact of the actions selected on the performance of the agent.*

The major weakness of the work at this stage is evaluation. The paper does not compare against any other similar learning algorithms or approaches, and does not assess performance against realistic opponents. It also considers only a single, highly stylized example game derived from a zero-sum game. A broader evaluating comparing against competing approaches is needed.

*We have expanded the evaluation part and added a new class of opponents. We compare performance to the existing algorithms RNR and SPS. While the example game is constructed to demonstrate the performance of RSRS, we have examined the performance in Battle of the Sexes, Chicken, a coordination game, a pursuit/evasion game, the Cuban Missile Crisis game, the Pursuit of the Israelites game, an inspector/worker game, a variant of Battle of the Sexes, a variant of Matching Pennies, Stag Hunt, and Traveller’s Dilemma, and found results consistent with those demonstrated in Rock/Spock/Paper/Lizard/Scissors.*

**READABILITY AND ORGANIZATION:**
It is reasonably organized but graphs are incredibly difficult to read when printed in black and white. Larger font and increased weight would be greatly appreciated.

*We have tried to make the graphs more readable. Unfortunately, space limitations prevented us from enlarging the figures significantly.*

**UPDATE:**
One issue is there seems to be a lot missing from the discussion of related work, and this work is not positioned very clearly in the literature. Papers on safety-level strategies, safe opponent exploitation, and related work on targeted optimality with safety in multi-agent learning (e.g., Chakraborty and Stone, ICML 2010) are all missing.

*We expanded the related work section and added the suggested citation.*

I am partly convinced by the arguments of another reviewer that the conceptual/theoretical contributions of the paper should be given more weight and have adjusted my scores accordingly, but I still view the empirical evaluation as a weak point for the paper.

– Summary: The paper introduces a new approach for balancing exploiting imperfect predictions of opponent play against safety in general-sum games, which is an interesting direction but the evaluation of the approach is currently lacking.

*We have extended the experimental part, addressing the concerns of the reviewer about the weakness of the empirical evaluation.*