We sincerely thank the reviewers for their careful reviews and thoughtful suggestions. We are in the process of incorporating many of the changes into the final version.

General comments: We appreciate the reviewers' comments regarding the value of our model in cyber-security and adversarial decision making environments in general. It appears that some of the reviewers' concerns arise from misunderstandings which we wish to address below:

(i) As commented by Reviewer #1, there is a concern regarding the "two-phase assumption." We would like to clarify that N-CIRL *does not* make this assumption; it is precisely what N-CIRL is trying to address. The two-phase assumption *only arises* in the IRL and MA-IRL environments where learning is done using *past* trajectories of the problem (termed demonstrations). The novelty of CIRL [Hadfield-Menell *et al.*, 2016] is the discovery that higher reward can be realized by an *online* method that intertwines learning and deployment. In the same spirit as CIRL, the proposed N-CIRL setting argues (and empirically demonstrates) that learning *online* from revealed information can lead to policies that yield a higher reward for the less-informed player (*i.e.*, the defender).

(ii) An additional clarification, relevant to both the first and second reviews, concerns the temporal nature of the unknown parameter. The term "non-stationary behavior" in line 73 refers to the non-stationarity *between* attacks, not *within* a given attack (where an attack is defined as a complete run of the game). We do not assume the attacker changes its intent as an attack is unfolding; as noted in footnote 3: "The intent parameter θ is further assumed to be fixed throughout the problem." The fact that the intent changes across different attacks is indeed the primary motivation for the development of N-CIRL. Namely, since the intent parameter θ may change between attacks, one cannot rely on previous attack data (which would almost certainly contain information for different attacks and attackers) to be informative for defending against new attacks. This is further emphasized by line 70 which states that the MA-IRL approach "is only useful if the goal(s) do not change between the learning and deployment phases". We apologize for the lack of clarity, and accordingly we will be removing any mention of non-stationarity in order to alleviate ambiguity.

(iii) As commented by Reviewer #2, there is a concern regarding the validity of the "information state reduction." Note that players are assumed to act *simultaneously* in each stage (as mentioned on line 104); neither the attacker nor defender is assumed to move first, that is, we *do not* model the (stage) game as a *Stackelberg* game. Additionally, based on the reviewer's comments, we discovered that we incorrectly wrote the range of the discount factor as the closed interval when it should be $[0, 1)$. The implication of the correction of this typo is that the game is finite and thus the value is guaranteed to exist. Under this setting, the correctness of the information state reduction can be obtained by [Sorin, 2003] (as shown by Lemma 1 in our paper), independently of [Rosenberg, 2000; Rosenberg and Vieille 2000]. Furthermore, due to the finiteness of the game, the issue regarding the non-existence of "maximin" in [Rosenberg and Vieille 2000] is no longer a concern as they consider the distinct (asymptotic) case where the discount factor approaches 1 (corresponding to an infinite game). Lastly, as a consequence of the reviewer's comments, the proof of the decomposition result can be more directly derived from [Mertens, Sorin, & Zamir, 1994. Repeated Games: Part A. *Université catholique de Louvain, CORE*]. We will reflect these changes in the final version.

Specific comments: We now individually address the remaining concerns made by the reviewers.

Reviewer #1: We hope your concerns have been addressed by the general comments above (specifically that N-CIRL does not assume two-phases, and restriction to one-stage strategies is, by Theorem 1, without loss of optimality).

Reviewer #2: We have addressed some of your concerns in the general comments, your remaining concerns are addressed below. First, our algorithm computes the value backup assuming a known model. Hence, we can use *linear programming* to reason about the attacker's strategies and compute the defender's strategy [Rosenberg, 1998. Duality and Markovian strategies. *Int J Game Theory*]. We agree that error analysis is needed for the model-free case (where one reasons about strategies via samples), but this is outside our paper's scope (as a first attempt to solve N-CIRL). Second, under the *information state reduction*, the sequential decomposition enables restriction to *one-stage* strategies without loss of optimality. Third, the fixed-point of the contraction map exists (since the discount factor is < 1) and is equal to the value of the game, with the resulting strategies forming a Nash equilibrium. Lastly, we will modify the text to incorporate your comments on the necessity of value alignment and the recency of IRL. We had made a typo about Θ; line 149 should be $\theta \in \Theta = \{1/3, 2/3\}$. Finally, we will include an algorithm sketch (via pseudocode) in the revision.

Reviewer #3: The contribution can be viewed by drawing an analogy to CIRL, which rests on theoretical results developed by [Nayyar *et al.*, 2013]. Similarly, we leverage existing theory to demonstrate that the proposed N-CIRL setting also has desirable structure, albeit, resulting in a harder problem. We are deriving complexity analysis results to formally show this. Regarding your other comments, we will be elaborating on the proof of Lemma 2 and moving it to the supplementary material (also enabling more of the related work to be in the main text). We agree with your concern regarding the illustration of the performance gain of N-CIRL over MA-IRL. Note that we allow MA-IRL to learn a *dictionary* of policies and test each one against an intent θ drawn from Θ. The plot illustrates the performance of N-CIRL on θ compared to *all* MA-IRL polices (thus MA-IRL has multiple points). Based on your comments, we realize that a fairer comparison is to compare the *average* performance of MA-IRL to *multiple* random θ. Furthermore, we will include a larger (less symmetric) setting that will more effectively illustrate the advantages of N-CIRL.