Machine learning techniques are increasingly being used to make consequential decisions in the real world in areas such as hiring (RagHAVAN et al. 2020). The rapid spread of machine learning and artificial intelligence has spurred an interest in researchers and regulators into ensuring these technologies are used ethically (Frankish and Ramsey 2014). A key theme of the ethical deployment of machine learning is that it takes into account and tries to minimize the harm done by its deployment. One approach to minimizing the harm done by machine learning, known as algorithmic recourse, recommends actions to stakeholders that, if taken, will reverse the output of the machine learning model that caused the harm.

Many different approaches to recourse have been proposed (Karimi et al. 2020) but this paper focuses on the one proposed by Karimi et al. (Karimi, Schölkopf, and Valera 2021) because it is both the state of the art, incorporates causal constraints, and is representative of the deficiencies of the single-agent approach to recourse.

Karimi et al. represent the situation of an agent seeking recourse from an entity as a structural causal model. A structural causal model \( M = (F, X, U) \) is a three tuple consisting of a set of endogenous variables \( X \), a set of exogenous variables \( U \), and a set of structural equations \( F = \{ f_1, ..., f_n \} \) that determine how values are assigned to the endogenous variables (Pearl 2009). There is a corresponding directed causal graph \( G_M \). In the graph, there is a node for every exogenous and endogenous variable. An edge connects a node \( n_i \) to \( n_j \) if and only if the variable corresponding to node \( n_i \) appears as an argument in the function \( f_j \) which assigns values to the variable corresponding to \( n_j \).

Counterfactual statements can be represented using do-operations of the form \( \text{do}(X_i := k) \). A do-operation transforms \( M \) to \( M_A \) by assigning \( k \) to \( X_i \) instead of assigning it the value of \( f_i \). The corresponding causal graph \( G_{M_A} \) is identical to \( G \) but with the incoming edges to the node corresponding to \( X_i \) removed. Recommended actions can be interpreted as a set of do-interventions \( A = \text{do}(\{ x_i := a_i \}_{i=1}^n) \) where \( I \) is the set of variables to be intervened on and \( a_i = x_i + \delta_i \). Assuming no hidden confounders and an invertible \( F \), any \( X \) can be uniquely determined given any \( U \) and vice versa. Hence any structural counterfactual query can be computed in the following way \( x^{\text{SCF}} = F_A(F^{-1}(x^F)) \).

Karimi et al. treat recourse recommendations as the solution to optimization problem (1).

\[
A^* \in \arg\min_A \text{cost}(A; x^F) \quad (1)
\]

s.t.

\[
h(x^{\text{SCF}}) \neq h(x^F),
\]

\[
x^{\text{SCF}} = F_A(F^{-1}(x^F)),
\]

\[
x^{\text{SCF}} \in P,
\]

\[
A \in F,
\]

The recourse interpretation of optimization problem (1) is an entity trying to recommend an action, \( A \), to an agent with feature values, \( x^F \), that minimizes the agent’s cost function while meeting a set of constraints that ensure action, \( A \), changes the feature vector to one that causes the binary model, \( h \), to change its output. The new feature vector, \( x^{\text{SCF}} \), is called the structural counterfactual. \( P \) is the set of plausible features that the agent could reasonably have and \( F \) is the set of feasible actions for an agent with features \( x^F \).

Karimi et al. assume a single-agent perspective where the recommended action is optimized with respect to only that agent. This is true of algorithmic recourse approaches generally (Karimi et al. 2020). In real decision-making contexts, the assumption of a single-agent environment does not hold. Decision makers will often be using machine learning to advise multiple agents in a context where an agent’s outcome...
is not independent of another agent’s actions. Therefore, recourse recommendations that are optimal with respect to a single agent may be sub-optimal when considered from a multi-agent perspective. For example, companies such as Apple, Google, and Waze offer GPS services that make route recommendations to many drivers simultaneously. The optimal route for a driver to take from point A to point B is in part a function of the expectation of the number of other drivers taking that route. It is established that certain network configurations can result in a situation whereby individual drivers taking the fastest route makes all of the drivers worse off (Nisan et al. 2007). Therefore, GPS recommendations designed to make a driver better of will make the group of drivers they advise worse off. There is evidence that this phenomenon is currently contributing to traffic congestion in the real world (Cabannes et al. 2018).

**Single-Agent vs Multi-Agent Recourse**

The following example, based on the classic strategic game the prisoner’s dilemma, is illustrative of the problems that can arise when using the single-agent perspective when there are multiple interacting agents. Assume an entity is advising two agents on how to reduce their current prison sentence. Each agent has the option to collaborate with the police and betray the other agent or remain silent and refuse to collaborate. The sentence reductions for each agent are a function of both their own action and the action of the other agent and are given by Table 1. Assume the entity has models $h_1$ and $h_2$ that perfectly predict the payoffs given in the table below for Agent 1 and Agent 2 respectively.

| Agent2 | Betray (1) | Silent (0) |
|--------|-----------|-----------|
| Betray (1) | 3.5, 3.5 | 10, 1     |
| Silent (0) | 1, 10    | 5, 5      |

The prisoner’s dilemma example can be stated as the following structural causal model. $M = \langle F, X, U \rangle$ where $X = \{ h_1, h_2 \}$, $U = \{ x_1, x_2 \}$, $F = \{ f_1, f_2 \}$, and

$$f_i(x_i, x_j) = \begin{cases} 1 & \text{if } x_i = 0, x_j = 1 \\ 3.5 & \text{if } x_i = 1, x_j = 1 \\ 5 & \text{if } x_i = 0, x_j = 0 \\ 10 & \text{if } x_i = 1, x_j = 0 \end{cases}$$

The corresponding causal graph $G_M$ is depicted in Figure 1.

Assume that Agent 1 comes to the entity and notes its current sentence reduction is given by $h_1(x^F) = [0, 1]$ year. $P = \{ [0, 0], [1, 0], [0, 1], [1, 1] \}$ and $F = \{ [1, 0], [0, 0] \}$. In this example, the optimal recourse according to the framework given by (1) is $a = x^F + \delta^* = [0, 1] + [1, 0] = [1, 1]$. The new predicted payoff for Agent 1 is $h_1([1, 1]) = 3.5$. The goal of recourse for that agent has been achieved. There are two additional facts to note. $h_2$ has gone from 10 to 3.5 on the new input $x^{SCF} = [1, 1]$. Furthermore, $h_1(x^{SCF}) + h_2(x^{SCF}) = 7 < h_1(x^F) + h_2(x^F) = 11$. The recourse recommendation has made Agent 2 worse off and decreased the sum of benefits to both agents.

Now consider the alternate case where Agent 1 comes to the entity and notes its current sentence reduction is given by $h_1(x^F) = [0, 0]$ = 5. The recourse recommendation that improves Agent 1’s outcome is $a = x^F + \delta^* = [0, 0] + [1, 0] = [1, 0]$. As before, this recommendation makes Agent 2 worse off. Unlike the previous case, the amount it makes Agent 1 better off is greater than the amount it makes Agent 2 worse off, so the group is better off.

The above examples illustrates inherent trade-offs that must be considered when making recourse recommendations in strategic situations where an agent’s outcome depends on her actions and those of other agents. An ethically ideal recommended action would (1) improve the prediction of the agent being advised (principal agent), (2) not worsen the prediction for any other agent, and (3) increase the sum of the predictions for all of the agents. The latter two properties are known in game theory as Pareto efficiency and social-welfare efficiency (Tadelis 2013). These are desirable properties when making recourse recommendations because they measure the social harm and robustness of the recommendation. Robustness is the ability of a recourse to be honored if acted upon and has been identified as a desirable property in the literature (Venkatasubramanian and Alfano 2020; König, Freiesleben, and Grosse-Wentrup 2021). If an entity makes a recourse recommendation that negatively affects the model output with respect to an agent it previously made a recommendation to, recourse is violated.

When considering the single-agent environment, any action that satisfies one property must trivially satisfy all three. However, when there are at least two agents, an action that satisfies one property can fail to satisfy one or both of the other two.

Our work is not the first to examine recourse from a multi-agent perspective. (Rawal and Lakkaraju 2020) propose a framework for examining the effects of recourse from a group perspective, but they do not incorporate causal
relationships between features when making recommendations. Furthermore, although they take into account that recourse recommendations might perform differently for different subgroups, they don’t address the idea that an agent acting on a recommendation might affect another agent’s model outcome.

Problem Reformulation
Below we propose extensions of the Karimi et al. framework to include multi-agent constraints.

Social-Welfare-Efficient Recourse
In the case that the entity wants to recommend an action that increases the sum of the predictions for the N-agents it advises, the following is a proposed revision of the optimization problem.

\[ A^* \in \arg \min_A \text{cost}_i(A; x^F) \]  \hspace{1cm} (2)
\[ \text{s.t. } \sum_{j=1}^{N} h_j(x^{SCF}) \geq \sum_{j=1}^{N} h_j(x^F), \]
\[ x^{SCF} = F_A(F^{-1}(x^F)), \]
\[ x^{SCF} \in P_i, \]
\[ A \in F_i, \]

If the entity is interested in making recourse recommendations that don’t decrease social welfare as opposed to increase it, then the inequality constraint can be made non-strict.

Pareto-Efficient Recourse
In the case that the entity wants to recommend an action that makes none of the N agents the entity advises worse off, the following is a proposed revision of the optimization problem.

\[ A^* \in \arg \min_A \text{cost}_i(A; x^F) \]  \hspace{1cm} (3)
\[ \text{s.t. } \forall j \in N, h_j(x^{SCF}) \geq h_j(x^F), \]
\[ x^{SCF} = F_A(F^{-1}(x^F)), \]
\[ x^{SCF} \in P_i, \]
\[ A \in F_i, \]

The inequality constraints from 1-3 can be combined. For example, an institution might find it ethically desirable to recommend actions that make Agent i strictly better off but make no other agent worse off or the group no worse off.

There are currently two major challenges towards solving problems such as (2) and (3). First, estimating causal relationships between features when making recommendations. Second, if the number of agents and features per agent is large, the resulting constrained non-linear optimization could be substantially more difficulty to solve than its single-agent analog.

Experimental Results
The prisoner’s dilemma is a common real-world multi-agent decision-making problem and appears in areas as diverse as arms control (Majeski 1984) and legal representation (Ashenfelter, Bloom, and Dahl 2013). Therefore, it is an ideal test case for demonstrating negative societal effects resulting from making single-agent recourse recommendations in a multi-agent environment. We tested the effects of our multi-agent optimization using experimental data from (Dal Bó 2005).

Dal Bó ran an experiment testing players’ strategies in a simulation of an iterated prisoner’s dilemma with infinite rounds. The iterated prisoner’s dilemma is a multi-round variant of the prisoner’s dilemma whereby a player’s outcome is the sum of their outcomes in the individual rounds of the game. In his experiment, he split players into a test and control group. The test group, designed to simulate infinite play, had each game consist of an initial round and then an unknown number of additional rounds where each additional round had a probability \( \rho \) of occurring. \( \rho \) was set to 0, 1/2, or 3/4. The corresponding control group players played a fixed number of games. The fixed number was set to either 1 game if it was a control for \( \rho = 0 \), 2 games for \( \rho = 1/2 \), or 4 games for \( \rho = 3/4 \), thus ensuring the control and test group played the same number of games on average. Test games with \( \rho \) equal to 0 and control games with one round are equivalent to the single round prisoner’s dilemma from Section 2. 3294 games met one of these criterion. Participants were paid an amount proportional to the number of the points they won in the game. Two different pay-off matrices were used during the game and are displayed below.

| Agent 1 | Agent 2 |
|---------|---------|
| Betray (1) | Silent (0) |
| Betray (1) | 35, 35 | 100, 10 |
| Silent (0) | 10, 100 | 65, 65 |

Because it is a simple controlled experiment, we know the structural equations a priori and can compute them for both the single-agent and multi-agent recourse recommendations from 1-3 using brute force. We examined the effects of recommendations made using single-agent optimization constraints. Of the recommendations made for the 3294 single-
round prisoner’s dilemma games, 434 recommendations resulted in an improvement for the principal agent as well as both a loss in social welfare and a violation of Pareto efficiency. No recommendations resulted in either social welfare or the opposing player improving. When either Pareto efficient or social welfare efficient constraints were added or just Pareto efficient constraints were considered alone, no recommendations were made at all because no action can improve the principal agent while improving social welfare or not harming the other player. Therefore, optimization setups from our framework would have prevented significant third-party harm when compared with single-agent recourse recommendations.

When recommendations were made with just social-welfare constraints, 2860 recommendations resulted in both an increase in social welfare and a decrease in the principal agent’s welfare. None of the recommendations resulted in an increase in the principal agent’s welfare. This result highlights a pressing ethical dilemma that previous work using the single-agent framework obscured, whether or not to advise the principal agent to do something that makes it worse off because it makes the group better off. In the literature, it has been suggested that explanations should never be made that encourage an action that would be harmful to the agent (Barocas, Selbst, and Raghavan 2020). Our analysis shows this approach can only be pursued at the cost of societal harm. It is important that the artificial intelligence community develops a consensus on how to handle ethical dilemmas such as this.

Table 3: Second payoff matrix used in experiments from (Dal Bó 2005).

| Agent 1   | Betray (1) | Silent (0) |
|-----------|------------|------------|
| Betray (1)| 45, 45     | 100, 10    |
| Silent (0)| 10, 100    | 75, 75     |

Table 4: Experimental Results.

| Criterion            | Single | Social Welfare | Pareto |
|----------------------|--------|----------------|--------|
| Single               | 434    | 0              | 0      |
| Social Welfare       | 0      | 2860           | 0      |
| Pareto               | 0      | 0              | 0      |
| single + social      | 0      | 0              | 0      |
| single + pareto      | 0      | 0              | 0      |

References

Ashenfelter, O.; Bloom, D. E.; and Dahl, G. B. 2013. Lawyers as agents of the devil in a prisoners dilemma game. *Journal of Empirical Legal Studies* 10(3):399–423.

Barocas, S.; Selbst, A. D.; and Raghavan, M. 2020. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 89–89. ACM.

Cabannes, T.; Fighiera, V.; Ugirumurera, J.; Sundt, A.; and Bayen, A. 2018. The impact of gps-enabled shortest path routing on mobility: a game theoretic approach. Transportation Research Board 97th Annual Meeting.

Dal Bó, P. 2005. Cooperation under the shadow of the future: Experimental evidence from infinitely repeated games. *American Economic Review*.

Frankish, K., and Ramsey, W. 2014. *The Cambridge Handbook of Artificial Intelligence*. New York, NY: Cambridge University Press.

Heinze-Deml, C.; Maathuis, M. H.; and Meinshausen, N. 2018. Causal structure learning. *Annual Review of Statistics and Its Application* 5(1):371–391.

Karimi, A.-H.; Barthe, G.; Schölkopf, B.; and Valera, I. 2020. A Survey of Algorithmic Recourse: Definitions, Formulations, Solutions, and Prospects. In *NeurIPS 2020 Workshop: ML Retrospectives, Surveys & Meta-Analyses (ML-RSA)*. Conference on Neural Information Processing Systems.

Karimi, A.-H.; Schölkopf, B.; and Valera, I. 2021. Algorithmic Recourse: From Counterfactual Explanations to Interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 353–362. New York City, New York: ACM.

König, G.; Freiesleben, T.; and Grosse-Wentrup, M. 2021. A causal perspective on meaningful and robust algorithmic recourse. In *International Conference on Machine Learning Workshop on Algorithmic Recourse*.

Majeski, S. J. 1984. Arms races as iterated prisoner’s dilemma games. *Mathematical social sciences* 7(3):253–266.

Nisan, N.; Roughgarden, T.; Tardos, E.; and . Vazirani, V. V. 2007. *Algorithmic Game Theory*. Cambridge University Press.

Pearl, J. 2009. Causality: Models, Reasoning, and Inference, 2.ed. New York, NY: Cambridge University Press.

Raghavan, M.; Barocas, S.; Kleinberg, J.; and Levy, K. 2020. Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 469–481. New York, NY: ACM.

Rawal, K., and Lakkaran, H. 2020. Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses. In Larchelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 12187–12198. Curran Associates, Inc.

Tadelis, S. 2013. *Game Theory: An Introduction*. Princeton University Press.

Venkatasubramanian, S., and Alfano, M. 2020. The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 284–293. ACM.