Available online at www.sciencedirect.com

ScienceDirect

Journal of Economic Theory 175 (2018) 1–37

JOURNAL OF Economic Theory

www.elsevier.com/locate/jet

Imperfect public monitoring with a fear of signal distortion

Vivek Bhattacharya [a,*], Lucas Manuelli [b], Ludwig Straub [c]

[a] *Department of Economics, Northwestern University, 2211 Campus Drive, Evanston, IL 60208, United States*
[b] *CSAIL, Massachusetts Institute of Technology, 32 Vassar Street, Cambridge, MA 02139, United States*
[c] *Department of Economics, Massachusetts Institute of Technology, 50 Memorial Drive, Cambridge, MA 02142, United States*

Received 25 July 2015; final version received 1 December 2017; accepted 9 January 2018
Available online 30 January 2018

Abstract

This paper proposes a model of signal distortion in a two-player game with imperfect public monitoring. We construct a tractable theoretical framework where each player has the opportunity to distort the true public signal and each player is uncertain about the distortion technologies available to the other player. We show that when players evaluate strategies according to their worst-case guarantees—i.e., are ambiguity averse over certain distributions in the environment—perceived continuation payoffs endogenously lie on a positively sloped line. We then provide examples showing that, counterintuitively, identifying deviators can be harmful in enforcing a strategy profile; moreover, we illustrate how the presence of such signal distortion can sustain cooperation when it is impossible in standard settings. We show that the main result and examples are robust to a number of natural modifications to our setting. Finally, we extend our model to a repeated game where our concept is a natural generalization of strongly symmetric equilibria. In this setting, we prove an anti-folk theorem, showing that payoffs under our equilibrium concept are under general conditions bounded away from efficiency.

JEL classification: C72; C73; D83

Keywords: Linearity; Imperfect public monitoring; Repeated games; Ambiguity aversion

* Corresponding author.
E-mail addresses: vivek.bhattacharya@northwestern.edu (V. Bhattacharya), manuelli@mit.edu (L. Manuelli), straub@mit.edu (L. Straub).

https://doi.org/10.1016/j.jet.2018.01.004
0022-0531/

1. Introduction

Many real-world strategic interactions are mediated by public signals that are possibly random functions of the players' actions, and economists have applied the theory of imperfect public monitoring to study many such situations. Applications include oligopoly games where price is influenced by quantity as well as demand fluctuations (see, for instance, Green and Porter, 1984), trade agreements with volatile trade volume (Bagwell and Staiger, 1990), and incentive contracts where workers' actions are unobserved (e.g., Radner, 1986 and Levin, 2003). In most of these settings as well as in the theoretical work on games with imperfect public monitoring (e.g., Abreu et al., 1990 and Fudenberg et al., 1994), it is taken for granted that the signal structure—the map from the action played to the public signal generated—is fixed and commonly known among all players. Recent papers (e.g., Fudenberg and Yamamoto, 2010, 2011) have acknowledged that this assumption is often especially strong and have proposed methods to relax it.

This paper proposes a new method to relax this assumption, based on the observation that in many of the applications above, players may fear that the signal that mediates their interaction can be *distorted* by their opponents. In a partnership game between two workers, say, compensation may be based on various dimensions of quality of an object that the workers produce jointly. Workers may then worry that their colleague may sabotage or otherwise alter the object after work on the project has concluded. In other settings, the signal is determined by a third party. Again in a worker-firm setting, promotions or bonuses may depend on performance evaluations conducted by a manager; a worker may worry about favoritism between his colleague and the manager that may cause the manager to doctor her evaluation in favor of the colleague. Cartel agreements are often based on measures like market share, which are computed by a consulting firm hired by the cartel.[1] Cartel members may worry that the consulting company is in the pocket of one of the firms and may be willing to alter these numbers in favor of this firm—perhaps in return for the promise of future business with this firm.

Signal distortion could directly be modeled as simply a game of imperfect public monitoring—in which players have a richer action space that allows them to affect signals without affecting per-period payoffs. Instead, we take a different and novel approach to modeling signal distortion in this paper. It is natural to believe that in many settings there is a large amount of uncertainty in how one's opponents can distort the signal as well as in how one will be able to distort the signal oneself. As such, in the model we present, we assume that players are unsure about the timing of the distortion as well as the distortion technology itself and are *ambiguity averse*, in the maxmin sense of Gilboa and Schmeidler (1989), over the possibilities. As a result, incentives are determined by "perceived" payoffs given by the preferences of ambiguity-averse agents. Another interpretation is that players choose actions that are "robust" to the elements of the environment over which they are uncertain and thus maximize their worst-case guarantees.

This setting is best described by a simple story involving a partnership game. Suppose two workers are both working on a project, and they can choose to either work hard or shirk. Each worker does not see what his colleague is doing, but their manager does see their actions and writes down performance evaluations about them. The manager will show these performance evaluations to her boss the following day to determine compensation for the workers. So far, the setting has exactly been one of imperfect public monitoring: the decision of whether to work

[1] One problem such third parties solve in a cartel, for instance, is that firms may be unwilling to share their books with competitors but may be willing to do so with a third party. Section 6.6 of Marshall and Marx (2012) discusses consulting firms and trade associations as potential third-party facilitators and provides many examples.

or shirk can be thought of as some game between the workers, the performance evaluation represents the signal generated by the actions, and the compensation represents the continuation payoffs. Now suppose that that evening, the workers can individually approach the manager and try to convince her to change the evaluation she will show her boss. When approaching the manager, each worker is unsure about (i) how the manager would be willing to modify the evaluation and about (ii) whether his colleague will be able to approach the manager later that evening. Given this uncertainty over the "distortion technology," we can imagine that when a worker approaches her, the manager simply offers him a take-it-or-leave-it offer to change the evaluation to something else in particular: because of ambiguity aversion over this distortion technology, worker 1 (say) would fear that the manager's offers to both workers would be especially undesirable to worker 1. We assume that the manager has no stake in this game and that there is no cost to convincing her to change the evaluation.[2]

The first main result of this paper is that the combination of the possibility of one's opponent distorting the signal and one being uncertain about and ambiguity averse over how one's opponent may modify the signal forces perceived continuation payoffs to lie on a positively sloped line. That is, the incentives provided by continuation payoffs are *perfectly aligned* across players, even if the true continuation payoffs are misaligned. This alignment occurs because players opt to align their own payoff with their opponent's in an effort to prevent their opponent from possibly distorting the signal to their disadvantage.

Our linearity result is reminiscent of a number of results in the literature on principal-agent models, such as Holmström and Milgrom (1987), Edmans and Gabaix (2011), Chassang (2013), and Carroll (2015). The intuition behind the linearity result in our model is most similar to that in Carroll (2015). Just as the principal in Carroll's model evaluates contracts according to worst-case guarantee over potential actions the agent can take, so do the ambiguity-averse players in our model evaluate continuation payoffs based on the worst-case over options that their opponents may be presented with.[3] As such, they hesitate to distort the signal to values that may harm their opponents, since they fear that their opponents would then be more willing to distort the signal to potentially worse outcomes. Consequently, the perceived payoffs are endogenously aligned. This key intuition does not rely on this specific extensive form but is rather robust to a variety of different models of the distortion phase, as we discuss in this paper. To our knowledge, ours is the first paper studying this sort of endogenous linearity in a game theory context rather than in contract design.

Our second main observation is that the equilibrium concept used in this paper can indeed have starkly different consequences from standard concepts studied in the literature. We present examples that suggest that, contrary to the intuition from standard games of imperfect public monitoring, making deviations by players *less* distinguishable can actually aid cooperation. Moreover, we suggest a reinterpretation of standard normal-form games and show that introducing a fear of signal distortion can *help* support Pareto-efficient outcomes, which would not be sustainable without this possibility of distortion. Finally, we extend our equilibrium concept to an

[2] We discuss at length the robustness of our results to various modeling assumptions in Sections 5 and 6.

[3] A subtle point, highlighted by a referee, is that Carroll (2015) does not impose ambiguity-averse preferences on the principal in his model. Rather, the principal acts *as if* she is ambiguity averse by judging contracts by their worst possible performance. One can interpret Carroll (2015) as an assumption on the procedure the principal uses to deal with uncertainty. Similarly, while we use the ambiguity aversion interpretation throughout the paper, it is possible to take the stance that players simply are unaware of the distribution of certain features of the environment and attempt to be robust to this uncertainty by using the procedure of evaluating outcomes by their worst-possible payoff.

infinitely repeated setting, where we show that it naturally generalizes the concept of a strongly symmetric equilibrium. Here, we prove an anti-folk theorem, namely that the set of possible equilibrium payoffs under our equilibrium concept is bounded away from efficiency.

We extensively discuss the robustness of our results to alternative setups and modeling assumptions, finding that across a variety of different setups, the combination of signal distortion and ambiguity aversion presents a powerful force towards alignment. Among other things, we study versions of our baseline model (i) where only a single agent has the opportunity to distort and (ii) where both agents have such an opportunity but there is no uncertainty about the timing. Even in case (i) with only a single agent distorting, perceived continuation values are shown to be "aligned" in the sense that the agents' continuation values are Pareto-ranked, albeit not necessarily linear. When letting the second agent distort as well in case (ii), despite having no uncertainty about the timing, perceived payoffs fall on a line again.

The assumption that players are ambiguity averse over distributions in the environment is nonstandard, but it is one in which there has been much recent interest in the context of mechanism design. Bose et al. (2006) study an auction in which there is uncertainty over the value distribution of the players, and Bodoh-Creed (2012) extends these results to more general ambiguity-averse preferences. Wolitzky (2016) considers trade between a buyer and seller who have maxmin beliefs over their opponents' types. Lopomo et al. (2010, 2011) consider principal-agent models (and more general mechanisms) in which Knightian uncertainty is embedded through incomplete preferences. Bose and Renou (2014) and Di Tillio et al. (2017) propose settings in which the mechanism designer is allowed to engineer ambiguity into an environment. Like these papers, we also modify a standard setting by introducing ambiguity to a particular feature of the environment and study differences induced by this ambiguity. We also relate to the literature on games with ambiguity-averse players, including Dow and da Costa Werlang (1994), Klibanoff (1996), and Lo (1996, 1999).

The remainder of the paper proceeds as follows. In Section 2, we present a baseline one-period model that formalizes our setup, explains the preferences we posit, and introduces the concept of *distortion equilibrium*—our key equilibrium concept in the paper. Section 3 proves the key result of the linearity of incentives and provides further discussion of the equilibrium concept and its robustness. We then provide some examples to motivate the intuition presented above—and draw distinctions between distortion equilibria and standard Nash equilibria—in Section 4. In Sections 5 and 6 we study the robustness of our results at length: Section 5 focuses on robustness with respect to alternative extensive forms of our baseline model, and Section 6 discusses a variety of other assumptions. Section 7 extends the model to an infinitely repeated game, proves our anti-folk theorem, and provides specific numerical examples to reinforce the intuition from our one-period examples in Section 4. Section 8 concludes. Appendix A contains an extension of both our baseline model and the variant we develop in Section 5 to N players. The Online Appendix collects a formal derivation of our equilibrium concept (Online Appendix B) and most of the derivations and proofs that characterize the infinitely repeated game (Online Appendix C), where we build on the work of Abreu et al. (1990), Fudenberg and Levine (1994), and Fudenberg et al. (1994) (henceforth APS, FL, and FLM, respectively).

2. A one-shot game

In this section, we consider a baseline model of signal distortion. While the setup in this section is specific to a two-player setting where each player gets one opportunity to distort the

signal, our main observations generalize to other extensive forms (described in Sections 5 and 6) and to $N > 2$ players (described in Appendix A).

2.1. Setup

Consider a two-player normal form game G where player i has a finite action set A_i. Payoffs are denoted by $g : A_1 \times A_2 \to \mathbb{R}^2$. An action profile $a = (a_1, a_2)$ also generates a public signal $y \in Y$ from a known distribution $\pi(a) \in \Delta Y$. We assume for simplicity that Y is finite. The realization y of the public signal will give player i an additional payoff $w_i(y)$. While we will assume that $w_i(y)$ is exogenously fixed, it can easily be thought of as the result of future strategic choices of the players; for instance, it can specify the equilibrium of a second-stage game that players coordinate on, or it can specify the future path of play in a repeated game of imperfect public monitoring.[4] We will thus refer to $w_i(y)$ as the *continuation payoff*. A (mixed) strategy for player i in this game is simply a distribution $\alpha_i \in \Delta A_i$, and the payoff to player i in this game from the mixed strategy profile $\alpha = (\alpha_1, \alpha_2)$ is[5]

$$\mathbb{E}_{\alpha_1[a_1],\alpha_2[a_2]}\left[(1-\beta)\cdot g_i(a_1,a_2) + \beta \cdot \mathbb{E}_{\pi(a_1,a_2)[y]} w_i(y)\right],$$

where $\beta \in [0, 1]$ is the discount factor. In what follows, we will use $g(\alpha)$ and $\pi(\alpha)$ as payoffs and distributions from mixed strategies in the obvious way.

In our setting, however, players will be given an opportunity to modify the signal before it is publicly revealed, thereby giving them the ability to modify the realized continuation payoffs. However, there is uncertainty as to whether one's opponent will be able to modify the signal after a player does so himself. Fig. 1 displays the extensive form game that models this uncertainty.[6] Players play actions α, and a signal $\hat{y}$ is drawn from $\pi(\alpha)$. However, the signal $\hat{y}$ is not immediately shown to the players; rather, there is a possibility that this signal $\hat{y}$ is distorted, and as such, we refer to this signal as a "temporary signal." After the temporary public signal $\hat{y}$ is drawn, Nature draws a signal distribution $\mu \sim F(\hat{y})$ for some distribution $F(\hat{y}) \in \Delta(\Delta Y)$. Note that this distribution can depend on the realization of the temporary signal $\hat{y}$. With probability γ, player 1 is selected to distort the signal first and is then given the choice between keeping $\hat{y}$ or accepting the alternate signal distribution μ.

- If player 1 decides to keep $\hat{y}$, then Nature draws another signal distribution $\mu' \sim F(a_2, \hat{y})$ (which can depend on the temporary signal $\hat{y}$ and the action player 2 played) and offers player 2 the choice between $\hat{y}$ and μ'. If player 2 decides on $\hat{y}$, then the true signal y becomes $\hat{y}$; otherwise, the true signal y is drawn from μ'.
- On the other hand, if player 1 decides to change to μ, then a second temporary signal $\hat{\hat{y}}$ is drawn from μ, and a distribution μ' is drawn from $F(a_2, \hat{\hat{y}})$ and presented to player 2, who can then choose between $\hat{\hat{y}}$ and μ'.

[4] See Section 7 and Online Appendix C for an application to repeated games.

[5] Throughout this paper, we will use the notation $\mathbb{E}_{\pi[y]} f(y)$ to denote the expectation of $f(y)$ where y is distributed according to π. That is, the dummy variable will be listed in square brackets in the expectation to make expressions easier to follow.

[6] See Section 5 for alternate extensive forms that also yield similar results, thus highlighting that the specific structure of Fig. 1 is not necessary for our main results.

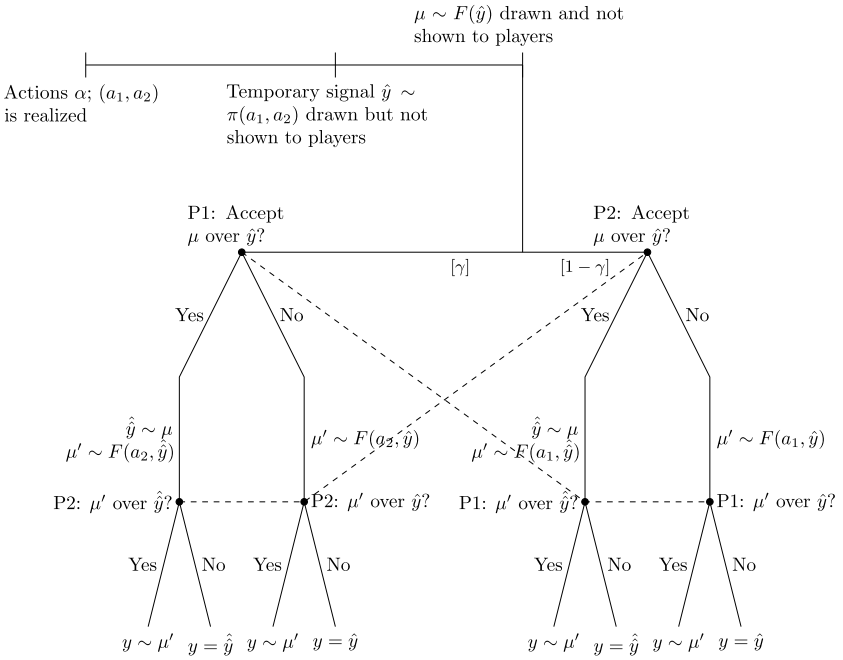

Fig. 1. An extensive form representation of the first-stage game. The tree begins in the top left. Probabilities for the moves of Nature are given in square brackets, and information nodes for the two players are connected by dashed lines. At the leaves of the tree, the signal y is announced as the public signal.

With probability $1-\gamma$, player 2 is selected to distort the signal first, in which case the same process applies with the roles of players 1 and 2 reversed.

Note that following the initial normal-form game G, nodes consist of tuples $(a_1, a_2, O, \hat{y}, \mu)$ in the middle row of Fig. 1 and tuples $(a_1, a_2, O, \hat{y}, \mu, YN, \hat{\hat{y}}, \mu')$ in the bottom row of Fig. 1. Here, (a_1, a_2) is the strategy realized in G, $O \in \{1, 2\}$ is the order of the players (i.e., which player modifies the signal first), $\hat{y}$ is the temporary signal shown to the first-moving player, μ is the alternate signal distribution shown to the first-moving player, $YN \in \{\text{Yes}, \text{No}\}$ is the decision of the first-moving player, $\hat{\hat{y}}$ is the temporary signal shown to the second-moving player, and μ' is the alternate signal distribution shown to the second-moving player. Player i's information set, however, is simply $h_i = (a_i, \hat{y}, \mu)$, as he does not know the order in which the players can modify the signal (i.e., whether his opponent will have a chance to modify the signal from what he chooses), nor does he know the action his opponent played.

Of course, the nonstandard aspect of this setup is the distortion phase, and it is worth briefly discussing the interpretation of $F(\cdot)$ and the alternate signal distributions μ drawn from this distribution. To do so, it helps to revisit the story from the Introduction. In this situation, we can interpret the signal y as a subjective evaluation by the manager of the agents, which is dependent on the true actions taken by the agents. The distribution $F(\cdot)$ can be interpreted as the distortion technology available to each player—or perhaps jointly to the pair consisting of a particular player and the manager together. For instance, it may be easier to change the signal y to a signal y' that is "close" in some unmodeled sense: perhaps the manager will not wish to alter her subjective evaluation too drastically. The realization μ from this distribution is the actual

distribution to which the manager (or the player) is able to change the signal when the player visits her.[7]

2.2. Distortion equilibrium

We now define main equilibrium concept in our paper under the assumptions alluded to in the Introduction. We assume that the economic agents are uncertain about *when* and *how* they will be able to distort the signal, and we model this uncertainty as ambiguity aversion over γ and the F's; that is, agents know that γ and F each belong to some set, but they are not sure about the actual values. More specifically, agents are risk-neutral with respect to probability distributions they know (i.e., outcomes of mixed strategies and the signal distribution) but are ambiguity averse over distributions they do not know.[8] We assume that agents have no information about γ other than the trivial bound that $\gamma \in [0, 1]$, and that agents believe each F is an element of $\Delta(\Delta(Y))^0$, the set of distributions over ΔY that have full support.

Online Appendix B derives the equilibrium concept in this section by starting with a sequential equilibrium of the game presented in Section 2.1 and formally introducing ambiguity aversion in γ and each F. Since the notation in that appendix is involved, we focus on the intuition here; footnote 9 connects the discussion in this section to that in the Appendix. When given the opportunity to distort the signal, the agent is ambiguity averse over the order in which the players are approached (γ) as well as the technology that will be available to their opponents if their opponents do indeed have a chance to re-distort the signal ($F(a_2, y)$). Once faced with a temporary signal, the worst possible situation for player 1 is that his opponent will have a chance to alter the decision, and that the technology available to player 2 will be such that he chooses the worst possible outcome for player 1. However, we do *not* exogenously assume that player 2's incentives are inherently misaligned from those of player 1: player 1 still only believes that player 2 will choose distortions that make player 2 better off (in player 2's own eyes, keeping in mind that player 2 also believes that he is distorting first). Therefore, when choosing actions in the first stage, player 1 effectively believes that (i) the alternate distribution he is offered will not allow him to improve the signal and (ii) his opponent will have the final say in distorting the signal.

We will describe the strategies in our equilibrium concept with ambiguity aversion with two objects: (i) an action in the game G followed by (ii) a *distortion strategy* for each player i, which we will denote $D_i : Y \rightrightarrows \Delta Y$. This distortion strategy must be a compact-valued correspondence with $\delta_y \in D_i(y)$ for each $y \in Y$, where δ_y is the Dirac distribution, placing probability 1 on y. $D_i(y)$ represents the set of distributions that player i is willing to accept in lieu of the temporary signal when the temporary signal is y.

[7] In our current setup, we have assumed that each player is offered a single alternative signal distribution to which he can choose to distort the temporary signal. We could imagine a more general model in which the players are allowed to choose from a *set* of possible distortion strategies instead of a single alternate signal. It is easy to see that such a model would not be operationally any different from the one proposed in this paper. As long as there is maximal ambiguity aversion over this set—i.e., players worry that any possible set of distributions could be offered to their opponents—it would be without loss of generality for each player to worry about singleton sets. Any non-singleton set could be replaced by the singleton set containing a (possibly unique) choice that a player would be willing to make from the larger set.

[8] We can derive the same equilibrium concept under the assumption that agents are infinitely risk-averse *only* with respect to the outcomes of the distributions governed by γ and F. Furthermore, by interpreting the payoffs as von Neumann–Morgenstern utilities this "risk neutrality" over all other distributions includes the cases of risk-averse and risk-seeking behavior.

Let $\mathcal{D}$ denote the set of all possible distortion strategies for either player, i.e., the set of all compact-valued correspondences D_i with $\delta_y \in D_i(y)$. Define the *perceived continuation value* from the signal y as

$$\tilde{w}_i(y) \equiv \min_{\mu \in D_{-i}(y)} \mathbb{E}_{\mu[y']} w_i(y'). \tag{1}$$

The interpretation is that player i perceives the continuation payoff from the signal y to be the one generated from the *worst possible distribution* that his opponent would prefer over y. Adopting this definition of perceived continuation payoffs, the time-zero utility of player i is defined to be

$$v_i(\alpha) \equiv (1-\beta) \cdot g_i(\alpha) + \beta \cdot \mathbb{E}_{\pi(\alpha)[y]} \tilde{w}_i(y), \tag{2}$$

with $\beta \in (0,1)$ the discount factor. To introduce the equilibrium concept, we define a notion of *consistency* between the perceived continuation payoffs in (1) and the distortion strategy. This condition essentially requires that the distortion strategy be optimal given the perceived continuation payoffs, i.e., that $D_i(y)$ is the set of distributions to which player i prefers to distort y. It is the analogue of a sequential rationality condition that would be imposed in a sequential equilibrium of the extensive-form game described in Section 2.1, keeping in mind that the agents are ambiguity-averse over the distributions from which their opponent's distortion opportunities will be drawn.[9]

Definition 1 *(Consistency).* A triple $(w, \tilde{w}, D)$ is said to be consistent if $\tilde{w}$ satisfies (1) and

$$D_i(y) = \{\mu \in \Delta(Y) : \mathbb{E}_{\mu[y']}[\tilde{w}_i(y')] \geq \tilde{w}_i(y)\}. \tag{3}$$

If $(w, \tilde{w}, D)$ is consistent, then $D_i(y)$ is clearly compact and contains δ_y, so the restriction that $D_i \in \mathcal{D}$ is implied by consistency. Implicit in (3) is a tie-breaking assumption that player i is willing to distort to any signal distribution that leaves him *weakly* better off.

Finally we define our equilibrium concept as follows.

Definition 2 *(Distortion equilibrium).* A strategy profile (α, D) is a distortion equilibrium given continuation payoffs w if

(i) defining $\tilde{w}(y)$ via (1), the triple $(w, \tilde{w}, D)$ is consistent as in Definition 1; and
(ii) for each player, α_i is optimal given α_{-i} and $\tilde{w}_i$, meaning for all $a_i \in \operatorname{supp} \alpha_i$,

$$a_i \in \arg\max_{a_i' \in A_i} \left\{ (1-\beta) \cdot g_i(a_i', \alpha_{-i}) + \beta \cdot \mathbb{E}_{\pi_y(a_i', \alpha_{-i})[y]} \tilde{w}_i(y) \right\}.$$

The main operational difference between the standard imperfect public monitoring setup and this alternate setup is that incentives in this setup are given by the *perceived* continuation values $\tilde{w}_i(y)$ instead of the standard continuation values $w_i(y)$.

We note that our notion of distortion equilibrium is closely related to that of a *multiple priors equilibrium* from Lo (1999). This equilibrium concept consists of a *set* of beliefs for each player, which collects all the beliefs over opponents' actions that the player "can imagine happening" and

[9] See Online Appendix B for details. Note that (1) is exactly analogous to $\tilde{U}_i(\delta_y|h)$ in (O.11) in Online Appendix B, and (2) is the analogue of (O.12). Consistency is exactly the formal counterpart of the way we construct the $D_i(y)$ sets in the sequential equilibrium, in (O.6).

over which he is ambiguity averse. It can be shown that any distortion equilibrium is equivalent to a multiple priors equilibrium applied to the extensive game described by Fig. 1 in which players are not ambiguity averse about each others' actions but only about some of nature's draws.

3. Properties of distortion equilibria

3.1. Linearity

Fix continuation payoffs $w(y)$. The consistency requirement for the D can be viewed as a fixed point problem between the distortion strategies and the perceived continuation payoffs $\tilde{w}(y)$. Given the $w(y)$ and a D, we can compute $\tilde{w}(y)$ as the solution to the minimization problem given in (1); given the $\tilde{w}(y)$, the $D(y)$ are uniquely defined from (3) in the definition of consistency. This fixed point procedure gives us the following key result.

Theorem 1. *Consistency of the triple $(w, \tilde{w}, D)$ requires that $\tilde{w}(y)$ lie on a line with slope in $(0, \infty)$, when plotting the pairs $\{(\tilde{w}_1(y), \tilde{w}_2(y))\}_{y \in Y}$.*

This proof is a consequence of two simple lemmas.

Lemma 1. *Suppose $\tilde{w}(y)$ and D satisfy consistency. Then, the $\tilde{w}(y)$ are strongly Pareto-ranked in that $\tilde{w}_1(y') \geq \tilde{w}_1(y'')$ if and only if $\tilde{w}_2(y') \geq \tilde{w}_2(y'')$.*

Proof. Suppose y and y' are such that $\tilde{w}_1(y) \leq \tilde{w}_1(y')$. Then, $D_1(y') \subseteq D_1(y)$. This implies that

$$\tilde{w}_2(y) = \min_{\mu \in D_1(y)} \mathbb{E}_{\mu[y'']} w_2(y'') \leq \min_{\mu \in D_1(y')} \mathbb{E}_{\mu[y'']} w_2(y'') = \tilde{w}_2(y'),$$

as needed. □

Lemma 2. *Suppose $\tilde{w}(y)$ and D satisfy consistency and, among the $\tilde{w}(y)$, we have a unique Pareto-best point $\tilde{w}_B$ such that $\tilde{w}_B \geq \tilde{w}(y)$ for all y and a unique Pareto-worst point $\tilde{w}_W$ such that $\tilde{w}_W \leq \tilde{w}(y)$ for all y.[10] Then, the $\tilde{w}(y)$ lie on a line with slope in $(0, \infty)$.*

Note, of course, that the fact that there is a unique Pareto-best point and a unique Pareto-worst point is implied by Lemma 1. Moreover, there may be multiple signals that give rise to these Pareto-extremal perceived continuation payoffs.

Proof of Lemma 2. First, if $\tilde{w}_{B,2} = \tilde{w}_{W,2}$ then Lemma 1 ensures that $\tilde{w}_B = \tilde{w}_W$, meaning $\tilde{w}_B = \tilde{w}_W = \tilde{w}(y)$ for all y. Trivially, all the $\tilde{w}(y)$ lie on a positively sloped line since they all coincide.

It remains to consider the case where $\tilde{w}_W < \tilde{w}_B$. Let $Y_B \equiv \{y \in Y : \tilde{w}(y) = \tilde{w}_B\}$ and $Y_W \equiv \{y \in Y : \tilde{w}(y) = \tilde{w}_W\}$. Let line ℓ connect $\tilde{w}_B$ to $\tilde{w}_W$, and suppose for contradiction that there exists $\hat{y}$ such that $\tilde{w}(\hat{y})$ lies above line ℓ.[11] First note that there exist y_B^* and y_W^* such

[10] When comparing vectors, we use $u \geq v$ to mean that $u_i \geq v_i$ for all components i.

[11] We are viewing the $\tilde{w}(y)$ as plotted on the $(\tilde{w}_1, \tilde{w}_2)$ plane. Also, the case where $\tilde{w}(\hat{y})$ lies below ℓ is symmetric; simply exchange the roles of players 1 and 2.

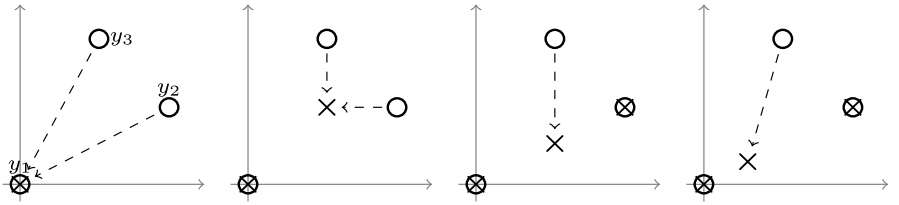

Fig. 2. Illustration of consistent triples. The circles represent $w(y)$, and the $\times$ represent corresponding $\tilde{w}(y)$ for various choices of $D_i(y)$. The $w(y)$ are mapped to the $\tilde{w}(y)$ via the dashed arrows.

that $w_2(y_B^*) = \tilde{w}_{B,2}$ and $w_2(y_W^*) = \tilde{w}_{W,2}$.[12] Note that $\tilde{w}_1(\hat{y}) = \alpha\tilde{w}_1(y_B^*) + (1-\alpha)\tilde{w}_1(y_W^*)$ for some $\alpha \in [0,1]$. It follows that $\alpha\delta_{y_B^*} + (1-\alpha)\delta_{y_W^*} \in D_1(\hat{y})$. But, $\tilde{w}_2(\hat{y}) > \alpha\tilde{w}_2(y_B^*) + (1-\alpha)\tilde{w}_2(y_W^*) = \alpha w_2(y_B^*) + (1-\alpha)w_2(y_W^*)$, which is a contradiction to the definition of $\tilde{w}_2(\hat{y})$ as $\min_{\mu\in D_1(\hat{y})} \mathbb{E}_{\mu[y]} w_2(y)$. □

It is important to stress that Theorem 1 is purely a result of the consistency requirement and is entirely independent of incentive compatibility in the actions chosen in G. Nevertheless, this theorem has strong implications for which strategies can be supported in an equilibrium. To explain this statement, note that the two ways to provide incentives that most theories offer are (i) value-burning and (ii) orthogonal enforcement.[13] The folk theorem in FLM relies crucially on the second method of providing incentives, as transfers of continuation values along the tangent hyperplanes are used to punish players if the public history is suggestive of a deviation from the prescribed strategy. In contrast, since the $\tilde{w}(y)$ are the values that give incentives, Theorem 1 implies that the incentives of the two players are perfectly aligned; that is, the only incentive provision mechanism is value burning. We can view the incentive provision mechanism in this setting as a generalization of the strongly symmetric equilibria of Green and Porter (1984) or Abreu et al. (1986): conditional on a choice of slope for the perceived continuation values, the incentives of the two players cannot be misaligned. However, unlike in strongly symmetric equilibria, the slope and the intercept between the incentives is a choice variable; this flexibility is relevant in Section 7 when extending this setting to an infinitely repeated game.

3.2. Existence and multiplicity

It is trivial to see that distortion equilibria always exist. Simply let α be a Nash equilibrium of the original game G, and let $D_i(y) = \Delta Y$ for all i and $y \in Y$. Then, $\tilde{w}(y) = \bigwedge\{w(y')\}_{y'\in Y}$ for all y,[14] and the continuation payoffs simply serve to shift the payoffs of the original game by a constant.

In general, however, distortion equilibria are not unique. Indeed, there are often infinitely many distortion equilibria for a particular game, as there are infinitely many consistent triples for a particular set of $w(y)$. Fig. 2 displays this multiplicity. We let Y be a three-element

[12] This is because $D_1(y_B) = \Delta Y_B$ for all $y_B \in Y_B$ and $D_1(y_W) = \Delta Y$ for all $y_W \in Y$, which implies that $\tilde{w}_{2,B} = \min\{w_2(y_B) : y_B \in Y_B\}$ and $\tilde{w}_{2,W} = \min\{w_2(y) : y \in Y\}$. Take the arg mins to find the desired signals y_B^* and y_W^*.

[13] See Sannikov and Skrzypacz (2010) for a discussion.

[14] The meet of two ordered pairs x and y, denoted $x \wedge y$, is the componentwise minimum. Denote the meet of a set S as meet $S \equiv \{x \wedge y : x, y \in S\}$. Finally, denote by $\bigwedge S$ the element x where $x_i = \min\{s_i : s \in S\}$; that is, $\bigwedge S$ is a tuple whose i^{th} coordinate is the minimum of the i^{th} coordinates of all elements in S.

set and fix $w(y)$ exogenously so that $w(y_1) = (0,0)$, $w(y_2) = (2,1)$, and $w(y_3) = (1,2)$. The points $(w_1(y), w_2(y))$ are plotted as circles. Each panel of Fig. 2 then shows a different set $\{(\tilde{w}_1(y), \tilde{w}_2(y))\}_y$, which can be rationalized with a different $D(\cdot)$. The first panel shows the case that we used to discuss existence: all $D_i(y) = \Delta Y$ and the perceived continuation payoffs collapse. The second panel illustrates the case where $D_i(y_1) = \Delta Y$ for both i, and $D_i(y) = \Delta(\{y_2, y_3\})$ for $y \in \{y_2, y_3\}$; this causes the "top two" continuation payoffs to collapse while there remains some distinction between different signals. In the third panel, $D_i(y_1) = \Delta Y$, $D_i(y_2) = \delta_{y_2}$, and $D_i(y_3) = \{(p_1, p_2, p_3) : 2p_2 + p_3 \geq 1\}$. This is the set of distributions such that the expected value of $\tilde{w}_i$ under that distribution is larger than $\tilde{w}_i(y_3)$. Finally, the fourth panel illustrates a consistent triple such that $D_i(y_1)$ and $D_i(y_2)$ are as before but the distortion set for y_3 is larger: in this case, $\tilde{w}(y_3) = (1/2, 1/4)$ and thus $D_i(y_3) = \{(p_1, p_2, p_3) : 2p_2 + p_3/2 \geq 1/2\}$.[15] It can easily be checked that these choices of D are also compatible with the definition of $\tilde{w}$ in (1).

The third and fourth panels of Fig. 2 show that once the positively sloped line on which the $\tilde{w}$ lie is chosen to be the one connecting $w(y_1)$ to $w(y_2)$, the location of $\tilde{w}(y_3)$ on this line is still indeterminate and can be picked as any point such that $\tilde{w}_i(y_3) \leq w_i(y_3)$ for both i. Note that if we were to posit $\tilde{w}(y_3) = (3/2, 3/4)$ so that $\tilde{w}_1(y_3) \geq w_1(y_3)$, then we would still be able to find D such that the pair $(\tilde{w}, D)$ still satisfies (3) from the definition of consistency. However, the triple $(w, \tilde{w}, D)$ would not be consistent, as $\tilde{w}$ would not be defined from (1); indeed, since $\delta_y \in D_{-i}(y)$ for all i, (1) requires that $\tilde{w}_i(y) \leq w_i(y)$. Perceived continuation values are necessarily more "pessimistic" than the actual continuation values.

The discussion suggests a very simple method for constructing consistent triples from exogenous $w(y)$. First, it is clear that the Pareto-worst $\tilde{w}_W$ will be $\bigwedge\{w(y)\}_{y \in Y}$. Next, we choose the Pareto-best $\tilde{w}$. To do so, we find a set $Y' \subseteq Y$ such that no element of $\{w(y)\}_{y \in Y'}$ is Pareto-dominated by some element of $\{w(y)\}_{y \notin Y'}$. By setting $D_i(y') = \Delta Y'$ for each i and $y' \in Y'$, we choose the Pareto-best $\tilde{w}_B$ as $\bigwedge\{w(y')\}_{y' \in Y'}$. In the second panel of Fig. 2, $Y' = \{y_2, y_3\}$ while in the third and fourth panels we have $Y' = \{y_2\}$. The consistency requirement affords a lot of flexibility in choosing $\tilde{w}(y)$ for $y \notin Y'$ such that $w(y) \geq \tilde{w}_W$ (for both components). *Any* point on the line connecting $\tilde{w}_W$ to $\tilde{w}_B$ such that $\tilde{w}_i(y) \leq w_i(y)$ for both i is a valid choice if the line on which the $\tilde{w}$ lie has slope in $(0, \infty)$.

The last sentence in the previous paragraph is not immediately obvious, and we now offer an explanation. Suppose we have a set of $\{\tilde{w}(y')\}$ that satisfy $\tilde{w}_i(y') \leq w_i(y')$ for each i and y' and also all lie on the line connecting $\tilde{w}_W$ to $\tilde{w}_B$. The choice of $\{\tilde{w}(y')\}$ pins down the $D_i(y')$ for each i and y', and thus we are interested in showing that the conjectured $(w, \tilde{w}, D)$ triple is indeed consistent, i.e., $\tilde{w}_i(y') = \min_{\mu \in D_{-i}(y')} w_i(\mu)$ for each i and y'. Since we assumed that the line connecting $\tilde{w}_W$ and $\tilde{w}_B$ has slope in $(0, \infty)$, we have $D_i = D_{-i}$ by Footnote 15. Fix a signal y. Thus, for any $\mu \in D_{-i}(y) = D_i(y)$, we will have

$$\mathbb{E}_{\mu[y']} w_i(y') \geq \mathbb{E}_{\mu[y']} \tilde{w}_i(y') \geq \tilde{w}_i(y),$$

since $D_i(y)$ is defined to be the set of μ such that the second inequality holds. Thus, it suffices to find a $\mu \in D_i(y)$ such that $\mathbb{E}_{\mu[y']} w_i(y') = \tilde{w}_i(y)$. As in Lemma 2, there must exist y_W and y_B such that $w_i(y_W) = \tilde{w}_{W,i}$ and $w_i(y_B) = \tilde{w}_{B,i}$. Furthermore, it must be that $\tilde{w}_i(y_W) \leq \tilde{w}_i(y) <$

[15] In any distortion equilibrium $D_1(y) = D_2(y)$ for all y. This is because we can write $\tilde{w}_1(y) = c\tilde{w}_2(y) + d$ for some $c \in (0, \infty)$ and d. Then,

$$D_1(y) = \{\mu : \mathbb{E}_\mu[\tilde{w}_1(y')] \geq \tilde{w}_1(y)\} = \{\mu : \mathbb{E}_\mu[c\tilde{w}_1(y')] + d \geq c\tilde{w}_1(y) + d\}$$
$$= \{\mu : \mathbb{E}_\mu[\tilde{w}_2(y')] \geq \tilde{w}_2(y)\} = D_2(y).$$

$\tilde{w}_i(y_B)$. Find the distribution μ^* that takes mass only on y_W and y_B such that $\mathbb{E}_{\mu^*[y']}\tilde{w}_i(y') = \tilde{w}_i(y)$. Then, $\mu^* \in D_i(y)$. But then $\mathbb{E}_{\mu^*[y']}w_i(y') = \tilde{w}_i(y)$ as well (since $\tilde{w}$ and w agree on the support of μ^*). Thus, the constructed $(w, \tilde{w}, D)$ satisfy (1) and are a consistent triple.

While there are often infinitely many consistent triples for a fixed w, we can define a natural refinement to restrict this set.

Definition 3 *(Minimally distortive).* Suppose $(w, \tilde{w}, D)$ and $(w, \tilde{w}', D')$ are both consistent triples. We say that $(w, \tilde{w}, D)$ is less distortive than $(w, \tilde{w}', D')$ if for any y $\tilde{w}(y) \geq \tilde{w}'(y)$ (entry by entry). We say $(w, \tilde{w}, D)$ is minimally distortive if there is no triple $(w, \tilde{w}', D')$ with $\tilde{w}' \neq \tilde{w}$ that is less distortive than $(w, \tilde{w}, D)$. A minimally distortive distortion equilibrium is one that is associated with a minimally distortive consistent triple.

The notion of "less distortive than" places a partial order on the set of consistent triples for a particular w. Intuitively, it provides a way to compare the extent of the fear of distortion. In Fig. 2, the first consistent triple—in which both players fear that their opponent will be willing to distort any signal to anything—is more distortive than the three other triples. The third triple is less distortive than the fourth. The second and third are not comparable to each other. Indeed, they are both minimally distortive. It is easy to see that there are finitely many minimally distortive consistent triples.[16] Furthermore, if there is a unique Pareto-efficient signal, then there is a unique minimally distortive consistent triple.[17]

There is an interesting interpretation of minimally distortive consistent triple in games with symmetric continuation values, that is, when the set $\{w(y)\} \subset \mathbb{R}^2$ is symmetric with respect to reflection around the 45-degree line. In every such game, there exists a unique minimally distortive consistent triple with perceived continuation values $\tilde{w}$ lying on the 45-degree line. The idea of continuation values falling on the 45-degree line is reminiscent of strongly symmetric equilibria in repeated games. Our result can be interpreted as a reason why incentives may be provided by continuation values on the 45-degree line *even if* actual continuation values that are not on the 45-degree line. Of course, if actual continuation values themselves already lie on the 45-degree line, perceived continuation values will do so, too. We discuss the connection to strongly symmetric equilibrium further in Section 7.

3.3. Discussion

Theorem 1 shows that in this model of signal distortion with ambiguity aversion, perceived continuation payoffs—and thus incentives—are perfectly aligned between two players in a game. The intuition is that each player is cognizant of the fact that distorting the signal in a manner that harms his opponent will only incentivize the opponent to alter the signal further (if he has the opportunity to do so). Given that each player is uncertain about what his opponent can do,

[16] Note that the construction of consistent triples discussed in this subsection consisted of two components: first, we pick a set $Y' \subseteq Y$ such that no element of $\{w(y)\}_{y \in Y'}$ is Pareto-dominated by some element of $\{w(y)\}_{y \notin Y'}$, and then we select the distortion strategies for $y \notin Y'$. Conditioning on Y', there is clearly only one way to select the distortion strategies for $y \notin Y'$ to ensure the triple is minimally distortive: simply choose $\tilde{w}(y)$ to be the point on the line joining the meet of $\{w(y)\}_{y \notin Y'}$ to the meet of $\{w(y)\}_{y \in Y'}$ that is closest to the latter point. Next, since there are only finitely many possible choices of Y', there are only finitely many minimally distortive consistent triples. Of course, not all set Y' lead to a valid minimally distortive consistent triple; the set Y' must be such that there is no point $y \in Y'$ such that $(w_1(y), w_2(y))$ is Pareto-superior to all other continuation payoffs in the set.

[17] In this case, there is only one valid choice of Y' that yields a minimally distortive consistent triple.

increasing the set of possibilities that the opponent is open to can only hurt the player. In this manner, each player chooses signals to effectively tie his (perceived) payoffs to his opponent's. This key intuition—that harming one's opponent only worsens his outside option and can thus make him more willing to harm others—is robust, as we discuss in Sections 5 and 6.

This result is reminiscent of a literature in contract theory that strives to explain linear incentive contracts designed by a principal for an agent. One early such paper is Holmström and Milgrom (1987), who present a model in which a contract that is linear in observable outcomes is optimal. The intuition, unlike in our model, is most closely related to the fact that such a scheme does not allow the agent to arbitrage nonlinearities in the incentive provision mechanism; it is in this sense that Holmström and Milgrom (1987) suggest that linear contracts are robust to a large strategy space. More recently, Edmans and Gabaix (2011) provides a separate set of sufficient conditions for linear contracts by considering a situation in which an agent decides on how much effort to exert after observing the noise in the system. Chassang (2013) considers a dynamic contracting environment and shows that linear contracts satisfy attractive properties, including performance bounds over many environments.

Carroll (2015) provides an explanation for linear contracts that contains many of the same ingredients that our model does. In his paper, an agent can choose a distribution of output at some cost, but the principal is ambiguity-averse over the set of distribution-cost pairs available to the agent. Linear contracts, in which the principal's payoff is tied directly to the agent's, then essentially guarantee that the principal also benefits from any self-interested action the agent takes. A similar setup and reasoning underlies our model as well. We can interpret the choice of signal distribution in our model as the analogue of the distribution of output available to the agent in Carroll (2015). There is an analogue of the cost of choosing such distributions as well: distributions that the player will not be able to choose (e.g., alternate distributions that will not be offered by the manager to the workers) may be prohibitively costly, and the distribution that player will be able to select can be thought of as costless. Each player, however, is uncertain over the distortions and costs available to his opponent. In equilibrium, even though there is no actual contract being designed by either player, payoffs are endogenously aligned.

The extensive form presented in Section 2 simply encapsulates three main components: (i) ambiguity aversion about the signal distortion technology of both oneself and one's opponent, (ii) ambiguity aversion about the timing of distortion, and (iii) uncertainty about the realized order at the time of distortion. We believe that these elements together capture a setting in which the players fear signal distortion at many stages of the game. In the first stage, while taking actions in G, (i) and (ii) together encapsulate that both players fear that their opponents will change the signal to something that happens to be unfavorable to them. The role of (iii) is to say that this fear of signal distortion exists *even at the time of distortion*, which is especially reasonable in many of the settings one may imagine. If a worker is worrying about possible favoritism between his colleague and the manager, it seems reasonable that a meeting with the manager still would not allay his worries. We relax (ii) and (iii) in Section 5 and show that the main results still hold.

We finally conclude with a short description of a tension in this model: there is uncertainty about the distortion technology available to the players, but players are expected utility maximizers conditional on a particular signal distribution. This tension is present in Carroll (2015) as well, and it has an especially clear interpretation in our model. In the settings we envision, there is usually an understood map between the signal—be it the evaluations of a manager, output from a team, or assessments from a third-party arbitrator—and the payoffs. The uncertainty we are modeling in signal distortion, however, may come from less quantifiable sources: a fear of favoritism, say, or the potential for sabotage. As mentioned in the Introduction, another inter-

pretation of our model would follow the terminology of "robustness" from Carroll (2015) and note that the decision procedure of the agents is to choose actions that are robust to whatever this unquantifiable uncertainty may be—by considering worst-case guarantees.

4. Examples

In this section, we present examples of games of imperfect public monitoring and compare the standard Nash equilibrium in the game to the distortion equilibria. The purpose of these examples is to highlight differences between these two forms of equilibria. Our first example underscores that improving the monitoring technology—so that deviations by individual players from a prescribed strategy are distinguishable—does *not* necessarily make it easier to sustain a certain first-stage action in equilibrium. Our second example notes that it is possible to sustain certain actions in a distortion equilibrium that are *not* sustainable in a Nash equilibrium.

In both examples, we investigate whether a certain action profile α is "sustainable" in the sense that it is possible to find suitable continuation values $w(y)$ in a certain (given) set W, and distortion strategies D such that (α, D) constitutes a distortion equilibrium.

4.1. *Identifying deviators can be* harmful

Consider the situation where G is a prisoner's dilemma, given as

G	C	D
C	$1,1$	$-1,2$
D	$2,-1$	$0,0$

.

There are three public signals, denoted y', y'', and y''', and we are choosing w from the set W to try to support the outcome (C,C) as an equilibrium. Suppose for concreteness that we are picking w from the convex hull of feasible outcomes in the prisoner's dilemma itself.[18] The distribution $\pi(\alpha)$ can be either

$S1$	y'	y''	y'''
(C,C)	$1/3$	$1/3$	$1/3$
(D,C)	0	$1/3$	$2/3$
(C,D)	0	$1/3$	$2/3$
(D,D)	0	0	1

or

$S2$	y'	y''	y'''
(C,C)	$1/3$	$1/3$	$1/3$
(D,C)	$2/3$	0	$1/3$
(C,D)	0	$2/3$	$1/3$
(D,D)	0	0	1

.

Under signal structure S1, cooperating yields an equal probability of each of the three signals. If either player deviates, then y''' becomes more likely. In this sense, it is possible to (statistically) see whether *some* player deviated but impossible to tell who it is. Under signal structure S2, the distribution induced by (C,C) remains the same as in S1. The difference between S1 and S2 is that in S2, it is possible to identify who deviated: if player 1 deviates, then y' is more likely, while y'' is more likely if player 2 deviates. Note that in both structures, if both players deviate,

[18] Choosing w from the set of feasible outcomes is reminiscent of w representing a true continuation payoff in a repeated game, which we explore in Section 7. In a different interpretation, we can imagine that the players play a second-stage game G' in which the Nash equilibria have payoffs $(1,1)$, $(-1,2)$, $(2,-1)$, and $(0,0)$, and a distortion equilibrium involves coordinating on a Nash equilibrium of G' after each signal.

then the signal will be y''' with certainty; however, since in this example we are trying to enforce (C, C) and are only concerned with unilateral deviations, this distribution is irrelevant.

First consider whether it is possible to sustain (C, C) as the outcome of standard Nash equilibria under signal structures S1 and S2. Under S1, the incentive compatibility condition for (C, C) for player 1 is that

$$(1-\beta)\cdot 1+\beta\cdot\left(\frac{1}{3}w_1(y')+\frac{1}{3}w_1(y'')+\frac{1}{3}w_1(y''')\right)$$
$$\geq (1-\beta)\cdot 2+\beta\cdot\left(\frac{1}{3}w_1(y'')+\frac{2}{3}w_1(y''')\right),$$

which simplifies to $w_1(y')-w_1(y''')\geq 3(1-\beta)/\beta$. Similarly, we find that the incentive compatibility condition for player 2 is that $w_2(y')-w_2(y''')\geq 3(1-\beta)/\beta$. Thus, for $\beta\geq 3/4$, one way to enforce (C, C) in the first stage of a distortion equilibrium is to set $w(y')=w(y'')=(1,1)$ and $w(y''')=(0,0)$. Under S2, the incentive compatibility conditions for players 1 and 2 are $w_1(y'')-w_1(y')\geq 3(1-\beta)/\beta$ and $w_2(y')-w_2(y'')\geq 3(1-\beta)/\beta$, respectively. In this case, with $w(y')=(-1,2)$, $w(y'')=(2,-1)$, and $w(y''')=(1,1)$, setting $\alpha=(C, C)$ is a Nash equilibrium with standard preferences as long as $\beta\geq 1/6$.

To compute distortion equilibria, we must specify a value $w(y)$ for each public signal y, compute a consistent triple $(w,\tilde{w}, D)$, and show that (C, C) can be sustained as a first-stage action in equilibrium given the perceived continuation values. Under S1, we have found $w_i(y)$ that already lie on a positively sloped line that sustain (C, C) as a Nash equilibrium. We can set $\tilde{w}_i(y)=w_i(y)$ and achieve consistency by setting $D_i(y')=D_i(y'')=\Delta\{y', y''\}$ and $D_i(y''')=\Delta Y$. However, note that the incentive compatibility conditions for (C, C) under S2 require that $\tilde{w}_1(y'')-\tilde{w}_1(y')\geq 3(1-\beta)/\beta$ and $\tilde{w}_2(y')-\tilde{w}_2(y'')\geq 3(1-\beta)/\beta$; these are the same as the Nash equilibrium, with w replaced by $\tilde{w}$. However, for any $\beta<1$, these conditions imply that $\tilde{w}(y')$ and $\tilde{w}(y'')$ line on a *negatively* sloped line, which contradicts Theorem 1. Indeed, under signal structure S2, it is impossible to sustain (C, C) in the first stage of a distortion equilibrium regardless of the continuation payoffs (i.e., even if they were not restricted to be in the convex hull of the prisoner's dilemma's payoffs).

In the above examples, moving from signal structure S1 to S2 facilitates cooperation in the standard case;[19] however, it actually hinders cooperation when studying distortion equilibria in that cooperation is *impossible* under S2. The difference between the two signal structures is that S1 is such that deviations by players are not distinguishable and result in the same signal distribution: whenever *either* player deviates from the prescribed strategy, there is a high likelihood of y'''. This does not allow for continuation payoffs where specifically the deviating player is punished. On the other hand, S2 does distinguish between a deviation by player 1 and one by player 2. A standard Nash equilibrium can leverage this distinction by using continuation payoffs of the "I win/you lose" form, thereby making both players unwilling to deviate and stomach these personal losses. Indeed, to satisfy the incentive compatibility conditions for (C, C) using S2, continuation payoffs *must* be in this anti-aligned form. This is not possible for perceived continuation payoffs, as shown by Theorem 1.

Informally, the fear of signal distortion induces player 1 to worry that player 2 will be able to "point the finger" at him by switching the signal to one that suggests that player 1 deviated. Un-

[19] We can see this informally by comparing the discount factors needed to support (C, C) in a Nash equilibrium. An alternate route would be compute something like the score in FL.

der S1, such finger-pointing is not possible, since there is no signal that suggests that a deviation by player 1 is any more likely than one by player 2. This is a key intuitive difference between this setup and the one in FLM, where distinguishing deviators is quite helpful in achieving enforceability. We should note, however, that the notion of "distinguishing deviators" used in this section is different from the pairwise full rank condition in FLM. Neither S1 nor S2 satisfies pairwise full rank for (C, C) since neither satisfies the pairwise-identifiability condition from FLM. However, S1 and S2 fail pairwise-identifiability for different reasons. Signal structure S1 fails since a convex combination of deviations from player 1 (i.e., playing D with probability 1 to generate the profile (D, C)) yields the same signal distribution as a convex combination of deviations from player 2 (also playing D with probability 1 to generate the profile (C, D)). Signal structure S2 does not fail because of this property: any deviation by player 1 can be statistically differentiated from any deviation by player 2.[20] Rather, it fails since a linear combination of the distributions from (C, C), (D, C), and (C, D) equals zero. Thus, in our example, our notion of being unable to distinguish deviators entails more than simply a failure of pairwise-identifiability.[21] We leave for future work a formal analysis of general conditions that enable us to rank signal structures based on whether is they make it more difficult to identify deviators.[22]

4.2. *Cooperating can be* easier

In this section, we again study the prisoner's dilemma, but instead of changing the signal structure between two games, we change G itself. To motivate this section, consider a variation of the story of the two workers given in the Introduction. Two prisoners are asked to play the prisoner's dilemma, but instead of reporting their decisions to a judge who immediately carries out the punishment, they report their decisions to a bailiff. The bailiff then relays the prisoners' decisions to the judge. However, the prisoners worry that there is a chance that the bailiff stops by one or both of their cells and gives them a chance to change the report he will present the judge.

To formalize this story, let G be a game with two actions for each player, C for cooperate and D for defect. Suppose that $g_i(a) = 0$ for all $a \in \{C, D\}^2 \equiv A^2$ and all i. There are four signals, denoted y_a for each $a \in A^2$. The signal structure is such that $\pi(a) = \delta_a$ for all $a \in A^2$. Moreover, $w_i(y_a) = \tilde{g}_i(a)$, where $\tilde{g}$ represents the payoffs from the standard prisoner's dilemma (e.g., the one from Section 4.1).

It is easy to see that the Nash equilibrium of this game is exactly the one in the prisoner's dilemma: both players play D. Indeed, the effective game in the first stage (taking into account payoffs from both G and the continuation payoffs w) is simply the same prisoner's dilemma, rescaled by β. However, (C, C) can be sustained as a (nontrivial) distortion equilib-

[20] It is impossible to find a convex combination of the distributions for (C, C) and (D, C) that equals a convex combination of the distributions form (C, C) and (C, D). We of course ignore the one that places mass 1 on (C, C), in which case neither player deviated.

[21] We should note, however, that failure of pairwise full rank for (C, C) is necessary for it to not be enforceable in a distortion equilibrium. Indeed, if a strategy profile α has pairwise full rank as in FLM, then it is always enforceable in a distortion equilibrium. This is a consequence of Lemma 5.4 in FLM: such a strategy profile is strongly enforceable (à la FLM) on all pairwise hyperplanes, including those that ensure that the continuation payoffs $w_1(y)$ and $w_2(y)$ lie on a line of strictly positive slope.

[22] One may hypothesize that, following Kandori (1992), Blackwell garbling would be one partial order over signal structures that make it "difficult to identify deviators" in our sense. However, note that S1 is not a garbling of S2, nor is it necessarily the case that garblings make it easier to sustain cooperation in a distortion equilibrium.

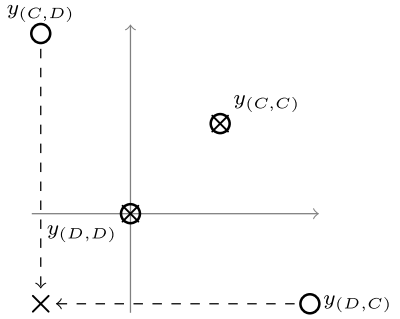

Fig. 3. Using a proxy in the prisoner's dilemma.

rium.[23] This construction is illustrated in Fig. 3. We have $D_i(y_{(C,C)}) = \delta_{y_{(C,C)}}$, $D_i(y) = \Delta Y$ for $y \in \{y_{(C,D)}, y_{(D,C)}\}$, and $D_i(y_{(D,D)}) = \left\{\left(p_{(C,C)}, p_{(C,D)}, p_{(D,C)}, p_{(D,D)}\right) : p_{(C,C)} - p_{(C,D)} - p_{(D,D)} \geq 0\right\}$, where p_a is the probability the signal distribution μ assigns to signal y_a. This induces a game with payoffs

$\tilde{w}$	C	D
C	$1, 1$	$-1, -1$,
D	$-1, -1$	$0, 0$

so that the payoffs of the entire game is simply β times the numbers in the above matrix. Thus, (C, C) can also be sustained as a distortion equilibrium. Note, however, that (D, D) can also be sustained.

The classic intuition behind the prisoner's dilemma is of course that if player 1 knows his opponent is cooperating, then it is a best response to defect. (That defecting is always a best response does not matter for our current purposes.) Suppose, however, the player 1 fears the bailiff will allow player 2 to alter the signal $y_{(D,C)}$, were player 1 to actually defect. The discussion in Section 2.2 suggests that player 1 would worry that player 2 would get the last word in distorting the signal. If player 2 were offered the temporary signal $y_{(D,C)}$, then he would be in an especially desperate situation and would be willing to alter the signal to anything—including $y_{(C,D)}$. Player 1 thus wants to avoid this and does not deviate to D in a distortion equilibrium. Note that player 2 would *not* change the signal $y_{(C,C)}$ to $y_{(C,D)}$—and player 1 realizes this—since he fears that player 1 will then be able to change the signal to something else.

The observation that introducing a "proxy"—an agent who, like the bailiff above, executes the action and can potentially be (costlessly) bribed by the players—into a game can help sustain efficient outcomes in a distortion equilibrium applies generally. In particular, consider a normal form game $\tilde{G}$ with action spaces A_i for $i = \{1, 2\}$ and payoffs $\tilde{g}_i(a) \in \mathbb{R}$ for $i \in \{1, 2\}$ and $a \in A_1 \times A_2$. The "proxied" version of this game is a game G with actions spaces A_i such that $g_i(a) = 0$ for all i and a, a signal set $Y = A_1 \times A_2$, and a signal structure $\pi(a) = \delta_a$. The Nash equilibria of the proxied game coincide with the Nash equilibria of the original game, but the distortion equilibria are substantially different. As noted in Footnote 23, any first-stage action profile can be trivially sustained as part of a distortion equilibrium of the proxied game if the

[23] Every strategy profile in the first-stage game can be rationalized as a distortion equilibrium if $D_i(y) = \Delta Y$ for all i and y, which would make the continuation payoffs trivial.

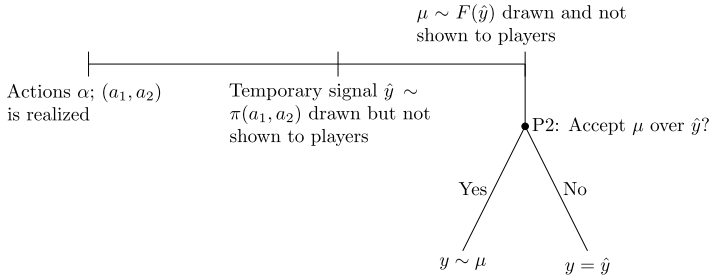

Fig. 4. An extensive form representation of a game in which only player 2 would be allowed to distort the signal.

agents expect an extreme level of distortion in the second stage ($D_i(y) = \Delta Y$). However, it is possible to devise reasonable refinements of distortion equilibria under which the Pareto-efficient action profiles are the unique action profiles which can be sustained as outcomes in a distortion equilibrium. For instance, as long as a minimally distortive equilibrium in a proxied game does not have all $\tilde{w}$ collapse to a single point, the only action profile that can be sustained in that equilibrium is a Pareto-efficient one.

5. Robustness to alternative extensive forms

The extensive form described in Section 2 captures a setting where two ambiguity-averse agents have the chance to distort a signal and there is no fixed timing of who goes first. It is a valid concern that alternative extensive forms may have dramatically different equilibrium properties, as illustrated by Ozdenoren and Peck (2008) in the context of a single agent playing a game against a malevolent nature. Interestingly, in this section we show that our main results are robust to a number of natural perturbations to the extensive form of the game. We extend our discussion of robustness to other aspects of our model in Section 6 below.

The two most important features of the extensive form in Section 2 are that (a) both agents get to distort the signal, rather than just one of them, and (b) agents are ambiguity averse about the order in which they get to distort the signal. We address both features in this section. In Section 5.1, we consider a very simple extensive form in which only a single agent distorts the signal. We show that this simple setup already implies that perceived continuation values $\tilde{w}(y)$ are Pareto-ranked. In Section 5.2 we then re-introduce the possibility that both agents distort the signal, but with common knowledge about the order in which agents are able to distort. In this case, we prove that perceived continuation values $\tilde{w}(y)$ are again on a line. The baseline model in Section 2 can be thought of as a symmetrized version of the model in which both players distort and have symmetric beliefs about the order.

5.1. A single player distorting

It is natural to ask how the results would change if players felt that the game were "rigged" in favor of a particular one of them. In this subsection, we model this alternative form by considering an extensive form in which only player 2 has the opportunity to distort the signal. Players are both still ambiguity-averse over the distortion technology F, but γ no longer plays a role. The setup is illustrated in Fig. 4: players play the stage game G, then player 2 receives an opportunity to distort the temporary signal $\hat{y}$ generated from $\pi(\cdot)$, and payoffs w are realized as a function of the distorted signal.

In this setting, the natural equilibrium concept is one in which player 2 distorts $\hat{y}$ to μ if $\mathbb{E}_{\mu[y]} w_2(y) \geq w_2(\hat{y})$ when the distortion stage occurs. When choosing actions in the game G, player 1 believes F is such that player 2 will distort to the worst-case signal for player 1, subject to his willingness to distort. In the first stage game, player 2 believes that the distortion stage will be such that player 2 will effectively not be able to distort. We can redefine consistency as follows, modifying Definition 1 to the case where just agent i distorts (here $i = 2$).

Definition 4 *(Consistency in Fig. 4).* A triple $(w, \tilde{w}, D_i)$ is consistent when player i distorts if

$$\begin{aligned} D_i(y) &= \left\{\mu \in \Delta(Y) : \mathbb{E}_{\mu[y']} w_i(y') \geq w_i(y)\right\}, \\ \tilde{w}_{-i}(y) &= \min_{\mu \in D_i(y)} \mathbb{E}_{\mu[y']} w_{-i}(y'), \text{ and} \\ \tilde{w}_i(y) &= w_i(y). \end{aligned}$$

As before, we call the $\tilde{w}$ the perceived continuation payoffs. To define a distortion equilibrium, we can modify Definition 2 as follows.

Definition 5 *(Distortion equilibrium in Fig. 4).* A strategy profile (α, D_2) is a distortion equilibrium, given continuation payoffs w, for the extensive form in Fig. 4 if

(i) $(w, \tilde{w}, D_2)$ is consistent as in Definition 4; and
(ii) for all $a_i \in \operatorname{supp} \alpha_i$,

$$a_i \in \arg\max_{a_i' \in A_i} \left\{(1-\beta) \cdot g_i(a_i', \alpha_{-i}) + \beta \cdot \mathbb{E}_{\pi(a_i', \alpha_{-i})[y]} \tilde{w}_i(y)\right\}.$$

It is easy to see that for a particular game, distortion equilibria exist. In this setting, consistency implies that the distortion strategy for player 2 is unique, as are the $\tilde{w}$.

What is the structure of perceived continuation values? Fig. 5 illustrates the transformation from w to $\tilde{w}$ in this setting. Panel (a) plots $(w_1(y), w_2(y))$ for a sample set of signals. Player 1 fears that player 2 will distort y_1 and y_2 to y_6. These maps, i.e., the arg min in Definition 4, are indicated by dashed arrows. Thus, player 1 sets $\tilde{w}_1(y_1) = \tilde{w}_1(y_2) = \tilde{w}_1(y_6) = w_1(y_6)$, for instance. Analogously, player 1 fears that y_3 would be distorted to a convex combination of y_5 and y_6; this convex combination is the minimum payoff player 1 can earn while keeping player 2 indifferent. Of course, player 2 does not believe that his perceived continuation payoffs are any different than the true continuation payoffs. Panel (b) plots the perceived continuation payoffs $(\tilde{w}_1(y), \tilde{w}_2(y))$ as solid black dots, with the original continuation payoffs as gray dots: points are moved horizontally relative to Panel (a), as shown by the dotted arrows.

The main observation in Panel (b) is that the perceived continuation payoffs are *Pareto-ranked*. The following theorem shows that this is a general property of continuation payoffs that satisfy consistency as in Definition 4, and the proof essentially extends the illustration in Fig. 5.

Theorem 2. *Let $w : Y \to \mathbb{R}^2$ and let $(w, \tilde{w}, D_2)$ be the unique associated consistent triple (see Definition 4). Then, $\tilde{w}$ are Pareto-ranked, i.e., if $\tilde{w}_1(y) > \tilde{w}_1(y')$ for any two signal realizations y, y', then $\tilde{w}_2(y) \geq \tilde{w}_2(y')$ (and vice versa if $\tilde{w}_2(y) > \tilde{w}_2(y')$). Moreover, the payoffs $\{\tilde{w}(y)\}_{y \in Y}$ lie on a concave arc: if for any y, y', y'' it holds that $\tilde{w}_1(y) < \tilde{w}_1(y') < \tilde{w}_1(y'')$, then*

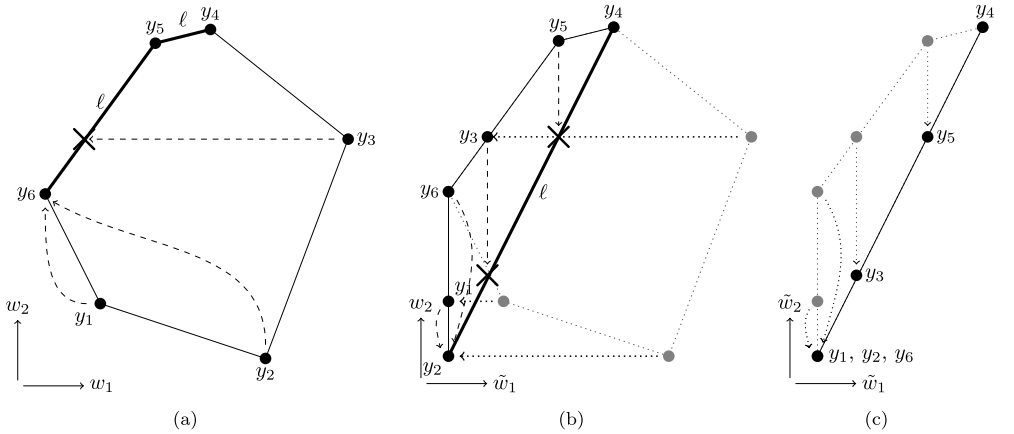

Fig. 5. Illustration of $(w, \tilde{w})$ for consistent triples as defined in Definitions 4 and 6. (a) The solid circles represent $w(y)$. The dashed arrows illustrate player 1's fears about how each signal will be distorted by player 2 (i.e., the worst-case signal distribution in $D_2(y)$) in the extensive form in Fig. 4. The thick solid line shows ℓ, which is defined in the proof of Theorem 2. (b) The gray circles represent $w(y)$ from (a). The dotted arrows illustrate how these payoffs were mapped to the perceived continuation payoffs $\tilde{w}(y)$, as defined in Definition 4 for the game in Fig. 4. These perceived continuation payoffs are denoted by black circles. The dashed arrows then illustrate player 2's fears of how player 1 would distort the signal in the extensive form in Fig. 6. (c) The solid circles the perceived continuation payoffs $\tilde{w}(y)$ for the game in Fig. 6, as defined in Definition 6. For reference, the payoffs $\hat{w}(y)$ (to use the notation from the proof of Theorem 3) are provided in gray, and the dotted arrows show the map from these payoffs to the perceived continuation payoffs.

$$\zeta \tilde{w}_2(y) + (1-\zeta)\tilde{w}_2(y'') \leq \tilde{w}_2(y')$$

where $\zeta \equiv \left(\tilde{w}_1(y'') - \tilde{w}_1(y')\right) / \left(\tilde{w}_1(y'') - \tilde{w}_1(y)\right) \in (0, 1)$.

Proof. Let $\underline{w}_1 \equiv \min_{y \in Y} w_1(y)$, and let $Y_W = \{y \in Y : w_1(y) = \underline{w}_1\}$ be the associated arg min. Let $\underline{y} \equiv \arg\max_{y \in Y_W} w_2(y)$; this is the best point for player 2 among the points that minimize player 1's continuation payoff.[24] Analogously define $\bar{y}$ to be the worst point for player 1 among points that maximize player 2's continuation payoffs. Let $B \equiv \text{bd co}\{(w_1(y), w_2(y)) : y \in Y\}$,[25] and consider the set

$$\ell \equiv \left\{(w_1, w_2) \in B : w_1(\underline{y}) \leq w_1 \leq w_1(\bar{y}) \text{ and } w_2(\underline{y}) \leq w_2 \leq w_2(\bar{y})\right\}.$$

It is easy to see that ℓ is a piecewise continuous collection of line segments with positive slope. We claim that for all y, either (i) $(\tilde{w}_1(y), \tilde{w}_2(y)) \in \ell$ or (ii) $\tilde{w}_1(y) = w_1(\underline{y})$ and $\tilde{w}_2(y) \leq w_2(\underline{y})$. The results then follow directly.

Let $Y_W \equiv \{y \in Y : w_2(y) \leq w_2(\underline{y})\}$. This is the set of signals y such that player 2 would be willing to distort to $\underline{y}$, i.e., that $\underline{y} \in D_2(y)$. Accordingly, $\tilde{w}_1(y) = w_1(\underline{y})$ for all $y \in Y_W$. Thus, for all $y \in Y_W$, (ii) holds. Now suppose $y \notin Y_W$. If $(\tilde{w}_1(y), \tilde{w}_2(y)) \notin \ell$, then there would be a point $(w_1, w_2) \in \ell$) with $w_2 = \tilde{w}_2(y)$ and $w_1 < \tilde{w}_1(y)$ which could be attained by some distribution μ' over signals. Since $w_2 = \tilde{w}_2(y) = w_2(y)$, $\mu \in D_2(y)$; this would contradict the condition that $\tilde{w}_1(y) = \min_{\mu} \mathbb{E}_{\mu[y]} w_1(y)$. Thus, it must be that $(\tilde{w}_1(y), \tilde{w}_2(y)) \in \ell$, which means (i) holds. □

[24] Assume that $\underline{y}$ is unique. This will only not be the case if two different signals have the same $(w_1(y), w_2(y))$, but arguments only undergo minor modifications in this case.

[25] We let co denote the convex hull of a set and bd denote the boundary.

What is the connection between Theorem 2 and the baseline linearity result in Theorem 1? The intuitions behind the two results are quite similar: if Player 2 is presented with a signal y with a low payoff $w_2(y)$, then he becomes more willing to distort this signal—and in the process may well harm player 1's payoffs as well. Harming player 2's "outside option" by choosing actions that yield low continuation payoffs for him is therefore not beneficial to player 1 either, just as choosing distortions that were harmful for player 2 would harm player 1 in the baseline setting. Indeed, the similarities between the two results become clear when presented with an alternate method to prove the Pareto-ranked result in Theorem 2. Suppose y' and y'' are such that $w_2(y') \leq w_2(y'')$. Then, $D_2(y') \supseteq D_2(y'')$, which means that $\tilde{w}_1(y') = \min_{\mu \in D_2(y')} \mathbb{E}_{\mu[y]} w_1(y) \leq \min_{\mu \in D_2(y'')} \mathbb{E}_{\mu[y]} w_1(y) = \tilde{w}_1(y'')$. Coupled with the definition that $w_2(y) = \tilde{w}_2(y)$ for all y, we have the desired result. Note that this argument is almost identical to the one presented in Lemma 2.

Theorem 2 also shows that as player 2's outside option $w_2(y)$ becomes lower, player 1 is marginally less and less affected since below a certain threshold—a convexity result that will prove useful below.

Note that the example from Section 4.1 continues to apply in this model, since perceived payoffs still cannot be anti-aligned in this extensive form. The logic in Section 4.2 also applies, even though the specific example considered there is uninteresting in the context of this section: all perceived continuation values $\tilde{w}$ collapse to a single point when using the equilibrium concept from Theorem 2, so (C, C) is trivially an equilibrium. A simple modification however gets around this knife-edge issue. Instead of the simplifying assumption $g = 0$, assume for example that $g(C, D) = (0, 1.5)$, $g(D, C) = (1.5, 0)$ and $g(\alpha) = 0$ otherwise, while keeping the total payoff $g(\alpha) + \mathbb{E}_{\pi(\alpha)}[w]$ unchanged.[26] In this case, the game without distortions still only admits (D, D) as unique Nash equilibrium, while the proxied game using the equilibrium concept from Theorem 2 only admits (C, C) as unique distortion equilibrium. At the time of playing the *first-stage action*—rather than the distortion, as in the example in Section 4.2—player 1 realizes that playing D would make player 2 more willing to distort the signal.

5.2. Common knowledge about the order of distortions

In the baseline model in Section 2.1, both players have the opportunity to distort the signal. In this section, we extend the extensive form in Section 5.1 to allow both players a chance to distort. In particular, we assume that after the first-stage game G is played, player 1 can distort the temporary signal, and then player 2 can distort this distorted signal. The key difference to the model in Section 2.1, however, is that here there is no ambiguity aversion with respect to the order of the distortions; in fact the order is common knowledge among the two agents. Fig. 6 shows the extensive form of this game.

In this setup, the natural equilibrium concept is that player 2 distorts signals based on the continuation payoffs w_2, which gives rise to perceived continuation payoffs $\tilde{w}_1$ for player 1. These perceived continuation payoffs are what give player 1 the incentives to distort the signal. When playing the first-stage game, as before, both players fear the distortion technology F is such that that they will not get a chance to distort the signal, and that their opponent will be able to distort it to the worst-case scenario. The following definition of consistency captures this idea.

[26] We can imagine a setting in which $\beta = 1/2$ so that payoffs are effectively the sum of the payoffs in the first stage game and the payoffs from the continuation signals.

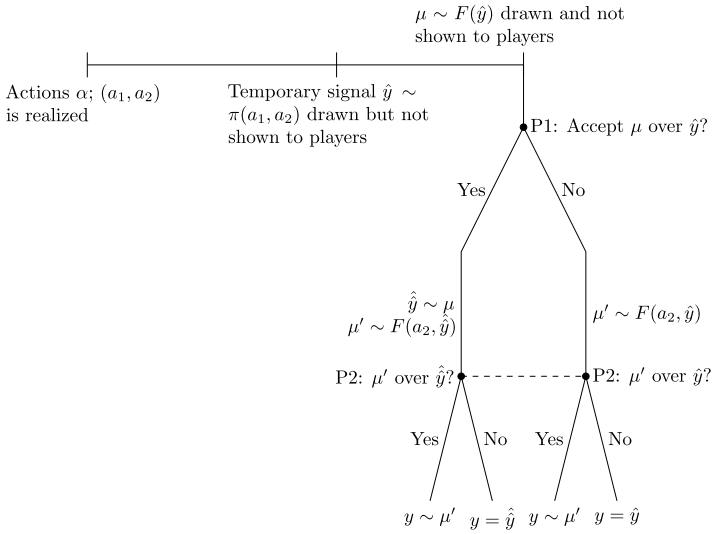

Fig. 6. An extensive form representation of a game in which the players distort the signal in a known order.

Definition 6 *(Consistency in Fig. 6).* A triple $(w, \tilde{w}, D)$ is consistent if

$$D_2(y) = \left\{\mu \in \Delta(Y) : \mathbb{E}_{\mu[y']} w_2(y') \geq w_2(y)\right\},$$
$$D_1(y) = \left\{\mu \in \Delta(Y) : \mathbb{E}_{\mu[y']} \tilde{w}_1(y') \geq \tilde{w}_1(y)\right\}, \text{ and}$$
$$\tilde{w}_i(y) = \min_{\mu \in D_{-i}(y)} \mathbb{E}_{\mu[y']} w_i(y').$$

The difference between Definition 6 and our original definition of consistency in Definition 1 is that under the extensive form in Fig. 6, player 2 distorts based on the actual continuation values w. Just as in Section 5.1, this implies that, given actual payoffs w, there are unique perceived payoffs $\tilde{w}$ and unique distortion sets D: D_2 is defined uniquely from w_2, which in turn defines $\tilde{w}_1$ uniquely and thus D_1 and $\tilde{w}_2$. Incorporating the first-stage game G, we can again define a distortion equilibrium exactly as in Definition 2, which only the consistency condition replaced.

With two agents distorting, we again find that perceived continuation values $\tilde{w}$ are perfectly aligned across the two players. Once again, the intuition behind this result is best captured through a diagram. Suppose the initial continuation payoffs are as in Fig. 5(a). Arguments from Section 5.1 show that before player 2 distorts, the perceived continuation values are given by the black dots in Fig. 5(b). To compute the perceived continuation values given that player 1 will also distort the signal, we can repeat the procedure that transforms Fig. 5(a) to (b). Note that player 2 fears that player 1—who distorts based on his perception of $\tilde{w}_1$—will be willing to accept y_2 in lieu of y_1 or y_6, and some combination of y_2 and y_4 in lieu of y_3 or y_5. These fears are indicated by dashed arrows, and they lead to continuation payoffs plotted in Fig. 5(c). The following theorem proves that the observation that these perceived continuation payoffs lie on a positively sloped line is more general than this example.

Theorem 3. *If $(w, \tilde{w}, D_2)$ are consistent as in Definition 6, then $(\tilde{w}_1(y), \tilde{w}_2(y))$ lie on a line with slope in $(0, \infty)$.*

Proof. Define the function $T^{(i)} : \{(w_1(y), w_2(y))\} \to \{(\tilde{w}_1(y), \tilde{w}_2(y))\}$ given by Definition 4 of consistency when player i distorts. Let $\{(\hat{w}_1(y), \hat{w}_2(y))\} \equiv T^{(2)}(\{(w_1(y), w_2(y))\})$ be the perceived continuation values before player 2 gets to distort. Then, note that under the definition of consistency given in Definition 6, the perceived continuation values $\{(\tilde{w}_1(y), \tilde{w}_2(y))\}$ can be expressed as $T^{(1)}\left(\{(\hat{w}_1(y), \hat{w}_2(y))\}\right)$. By Theorem 2, there exists a Pareto-worst signal $\underline{y}$ such that $\hat{w}_1(\underline{y}) \leq \hat{w}_1(y)$ for all y and $\hat{w}_2(\underline{y}) \leq \hat{w}_2(y)$ for all y. There also exists a Pareto-best signal $\bar{y}$ such that $\hat{w}_1(\bar{y}) \geq \hat{w}_1(y)$ for all y and $\hat{w}_2(\bar{y}) \geq \hat{w}_2(y)$ for all y. Note that $\underline{y}$ and $\bar{y}$ correspond to the same quantities as in the proof of Theorem 2.

By the same argument as in the proof of Theorem 2, $(\tilde{w}_1(y), \tilde{w}_2(y))$ will either (i) lie on the line segment ℓ connecting $\hat{w}(\underline{y})$ to $\hat{w}(\bar{y})$ or (ii) be such that $\tilde{w}_1(y) < \hat{w}_1(\underline{y})$ and $\tilde{w}_2(y) = \hat{w}_2(\underline{y})$. However, note that $\tilde{w}_1(y) = \hat{w}_1(y)$ for all y by the property of $T^{(i)}$, so (ii) is not a possibility here. It follows that all points lie on ℓ, which is a line with strictly positive (and finite) slope, due to the convexity result from Theorem 2.[27] □

The linearity result in this setting can be traced back to the convexity result presented in Theorem 2. Since $\tilde{w}_1$ is convex in the outside option $w_2(\hat{y})$ which player 1 presents to player 2, he is better off mixing between the best and worst possible outcomes. By engaging in this mixing, player 1 aligns his payoffs with player 2's, ensuring that player 2 can do no harm to him during player 2's distortion stage. It is this voluntary alignment out of fear of further signal distortion that lead to linearity both in our original setup of Section 2 and in the model of the current section. Note that in this extensive form, the examples in Section 4 go through just as in Section 5.1.

6. Further discussion of robustness

While Section 5 discusses how alternate extensive forms change the main result of the paper, we can also explore the effect of making other changes to the setup. In this section, we explore a number of such extensions informally; we include proofs and formalizations for some of these extensions in the appendix.

6.1. Multiple opportunities to distort

How important is the assumption that each player has at most a single opportunity to distort the signal? The specifics of the potential orders of distortions presented in Fig. 1 are not especially important. It is of course possible to extend the extensive form so that there are other distortion arrangements; for instance, player 1 could be allowed to distort once before player 2 is allowed to distort, and then player 1 could be allowed to distort again. Suppose that the information sets are such that at the time of distortion, each player believes that his opponent will have a chance to distort the signal again—i.e., the fear of future signal distortion persists throughout the entire game. Then, at the time of distortion, player 1 will believe that (i) at all of his own future distortion opportunities, F will be such that it is better to leave the signal undistorted and (ii) at his opponent's future distortion opportunities, player 2 will be presented with the worst-case opportunity for player 2. Then, Definition 1 will still be the natural one for consistency, and the structure of distortion equilibria will coincide with Section 3.

[27] Of course, it is still possible that $\tilde{w}(y)$ is the same for all y.

Similarly, the results we derived for our alternative extensive forms in Section 5 naturally carry over to a situations with multiple opportunities to distort. This is trivial for Section 5.1, where only a single player gets to distort. In Section 5.2 any sequence of distortion phases for players 1 and 2 lets perceived continuation values fall on a positive line. The reason is that the function $T^{(i)}$, which takes continuation values after a specific distortion phase by player i and converts them into perceived continuation values before the distortion phase, maps the set of (linearly) aligned continuation values into itself. Therefore, any sequence $T^{(i_1)} \cdots T^{(i_n)}$ with each player appearing at least once, must map any actual continuation values w into (linearly) aligned perceived continuation values $\tilde{w}$. We provide a formal argument in A.2.

6.2. $N > 2$ players

So far, the games we have studied only involve two players. We now discuss extensions to more players. In our baseline setup, modeling the distortion phase becomes more involved when considering extensions of this game to $N > 2$ players. Players may worry about which opponent will distort in the future, or perhaps about the order in which his opponents will distort, or even about whether his opponents will jointly distort the signal. Appendix A extends the concept of distortion equilibrium to games with N players and considers such issues. We show an analogue of Theorem 1 (see Theorem 7), directly extending the arguments in Section 3.1 under many reasonable models for the distortion phase in a game with three players. We can also extend the models in Section 5 to N players. Induction on N shows that if $N - 1$ players distort then perceived payoffs are Pareto-ranked; if all N distort, then they are linearly related among all players (see Theorem 8). Appendix A shows the formal results.

6.3. Other assumptions

Distortion technology Some of our assumptions are deliberately extreme to provide a clean analysis, but are immaterial to the intuition of the result. For instance, we have assumed that the players fear distortions to *any* signal distribution in ΔY. We can imagine instead a model in which players only fear distortions to signal distributions in some compact subset $S \subseteq \Delta Y$, which does not depend on the particular signal y being distorted. An example would be the case where S is the set of Dirac delta distributions on Y, so that agents only fear distortions to other signals (rather than signal distributions). In such a situation, perceived continuation values would still be strongly Pareto ranked, as Lemma 1 would still hold, but they would not necessarily lie on a positively sloped line. As the set S grows to approach ΔY, the perceived continuation values would become closer linear ones. In Section 5, allowing players to distort only to Dirac delta distributions would yield a Pareto-ranked result in both extensive forms, but the convexity result in Theorem 2 would not necessarily hold, nor would the linearity result in Theorem 3.

We have also assumed maxmin ambiguity aversion over F, following Gilboa and Schmeidler (1989). Analogous assumptions are rather common in this literature: Carroll (2015) assumes the principal is maxmin ambiguity-averse over the agent's technology, Di Tillio et al. (2017) consider an agent who is maxmin ambiguity-averse over the mechanism chosen by the seller, and Chung and Ely (2007) provide a foundation for dominant strategy mechanisms by studying a designer who has maxmin preferences over the agents' beliefs. These models are appealing since they can be interpreted as a having a "robustness" property to any possible belief of the environment. Compared to models of smooth ambiguity aversion, these models allow for very clean benchmarks. We would not usually expect such results to carry over one-for-one to models

with smooth ambiguity aversion, nor would we expect similar results when the technology in question—F in our case—is common knowledge and agents are at most risk-averse.

No ambiguity aversion over γ In our benchmark model in Section 2 we assume that players are not only ambiguity averse with respect to future distortion technologies F, but also with respect to the probability γ that Player 1 is the first agent to distort the signal. In Section 5.2, we consider a case where $\gamma = 1$ (the one with $\gamma = 0$ is analogous). We show that our results still go through in these polar cases for γ. Intermediate but known levels of γ turn out not to be quite as simple and do not necessarily have to be linear, since the perceived payoffs $\tilde{w}$ could be in a different order along the $\gamma = 0$ line than they are along the $\gamma = 1$ line. Of course, the model has a continuity property that as γ approaches 0 or 1, the perceived payoffs would approach ones that are linear.

Introducing a chance that there is no distortion phase In Sections 2 and 5, we assume that the distortion phase occurs with probability 1. We can enrich the model by introducing a probability ξ that the distortion phase actually occurs, so that the realized signal is simply the temporary signal $\hat{y}$ with some probability $1 - \xi$. If ξ is known, then the perceived continuation values are simply a convex combination of the actual continuation values w and the values $\tilde{w}$ discussed so far in this paper. As $\xi \to 1$, therefore, the perceived continuation values would approach linear ones.

We can also consider the setting when players are ambiguity averse over ξ, and they simply know that $\xi \in [0, 1]$. In this case, it is easy to see that in all extensive forms we have considered so far, both players would believe that $\xi = 1$; this is a consequence of the fact that $\tilde{w}_i \leq w_i$ in all situations, so players would always fear a distortion that either lets their opponent distort the signal (or at least leaves the signal unchanged).[28]

Costly distortions In addition, we have assumed no direct costs of distorting the signal, which may be appropriate for situations like favoritism but inapplicable to settings where significant bribes are required to persuade a third party to alter the signal. One way such costs can be introduced is by assuming that distorting a current signal y to some distribution μ to gain an expected utility benefit of $\Delta \equiv \mathbb{E}_\mu[w_i] - w_i(y)$ incurs costs $c_i(\Delta)$, where $c : \mathbb{R} \to \mathbb{R}_+$ is a non-negative cost function. With general cost functions, there is obviously no guarantee for our results to continue to hold. After all, very large costs get rid of distortions altogether.

However, progress can be made if costs are not prohibitively large. For example, suppose that distortions with benefits in some interval $\Delta \in (0, \overline{\Delta})$ are still worth doing, $\Delta > c(\Delta)$ for $\Delta \in (0, \overline{\Delta})$, and suppose that $\overline{\Delta}$ is not too small; in particular, assume $\overline{\Delta} > w_2(\underline{y}) - \min_y w_2(y)$, where $\underline{y}$ is defined as in the proof of Theorem 2. Then, even though many potential distortions maybe be costly, possibly prohibitively so, our main results in Section 5, Theorems 2 and 3 still go through. A similar condition on $\overline{\Delta}$ also ensures that our main result in Section 2, Theorem 1, holds despite costs.

7. An infinitely repeated game and an anti-folk theorem

Thus far, we have assumed that the continuation payoffs are exogenously fixed, and this simplification allowed us to cleanly describe much of the relevant intuition, along with the main

[28] Even in the extensive form in Fig. 4, we can say that player 2 believes $\xi = 1$, since he also believes that F will be such that he will effectively not be able to distort the signal at all. Behavior would be no different if he believed $\xi = 0$.

incentive alignment result in this paper (Theorem 1). It is nevertheless interesting to study methods to endogenize the continuation payoffs, and embedding the one-period setup in a supergame is one such method. In this section, we will define a *recursive distortion equilibrium*, a natural generalization of the equilibrium concept presented in the one-period model of Section 2 to a repeated game setting by first defining the concept recursively. We then show that under this solution concept, there is a natural anti-folk theorem in that perceived payoffs are bounded strictly away from efficiency. We also relate our solution concept to PPEs where continuation values have to lie on a positively sloped line, thereby establishing its property as a possible generalization of strongly symmetric equilibria. We finally show examples of comparative statics of Q_D with respect to the signal structure, mirroring our example in Section 4.1.

7.1. Recursive distortion equilibria

We use the same setup as in Section 2. Two players $i \in \{1, 2\}$ are playing a stage game G with action set A_i for player i and payoff matrix given by g. Moreover, there is a finite set Y of signals and a known map $\pi : \prod_i \Delta A_i \to \Delta Y$ that gives the signal structure as a function of actions taken in a given period. Players have a common discount factor β. A *public history* at time t is a sequence of all observed signals until time t; we let the set of all public histories be $\mathcal{Y}$ and let the initial history be denoted by $\emptyset$. We begin with a recursive definition of equilibrium that is the natural analogue to the distortion equilibrium defined in the previous section.

Definition 7 *(RDE).* A recursive distortion equilibrium (RDE) is a triple $\left(\alpha(y^{t-1}), D(y^{t-1}), w(y^{t-1})\right)$ of a (mixed) strategy $\alpha(y^{t-1}) \in \prod_i \Delta A_i$, a distortion strategy $D(y^{t-1}) \in \mathcal{D}^2$, and continuation payoffs $w(y^{t-1}) \in T$, for some bounded set $W \subseteq \mathbb{R}^2$, such that for each public history $y^{t-1} \in \mathcal{Y}$

(i) if we define

$$\tilde{w}_i(y^{t-1}, y_t) \equiv \min_{\mu \in D_{-i}(y^{t-1})(y_t)} \mathbb{E}_{\mu[y]}\left[w_i(y^{t-1}, y)\right],$$

then the triple $(w(y^{t-1}, \cdot), \tilde{w}(y^{t-1}, \cdot), D(y^{t-1}))$ is consistent as in Definition 1;

(ii) the Bellman equations

$$w_i(y^{t-1}) = \max_{a_i} \left\{(1-\beta) g_i(a_i, \alpha_{-i}(y^{t-1})) + \beta \mathbb{E}_{\pi(a_i, \alpha_{-i}(y^{t-1}))[y]}\left[\tilde{w}_i\left(y^{t-1}, y\right)\right]\right\},$$

are satisfied; and

(iii) $\alpha_i(y^{t-1})$ solves the previous maximization for player i.

We denote by $E_D(\beta)$ the set of RDE payoffs $w(\emptyset) \in W$ when the discount factor is β.

In the one-period model, note that we essentially specified an action, a distortion, and a vector of continuation payoffs at time 0 for each player. In this infinitely repeated model, we are specifying an action, a distortion, and a vector of continuation payoffs for each player *and each public history*; it is in this sense that recursive distortion equilibrium is a natural generalization of distortion equilibrium to an infinitely repeated game. The only additional constraint, of course, is that the continuation payoffs we specify at a particular history have to be compatible with those from different histories via the Bellman equation.

It is easy to see that as a result of the consistency condition, $\tilde{w}_i(y^{t-1}, \cdot)$ lie on a positively sloped line at all public histories. As such, incentives are again provided by (perceived) value-burning in this infinitely repeated game. An important distinction, however, between this refinement and equilibria in strongly symmetric strategies is that the slope and intercept of the perceived continuation payoffs are history-dependent choice variables, and this provides a method of differentially incentivizing different players. We develop the connection to SSEs further in Section 7.3.

For the purposes of comparison to FLM, a more relevant solution concept would be one in which strategies bear direct resemblance to public perfect equilibria. In Online Appendix C, we formally define such a *public perfect equilibrium with distortion* (PPED) and show in Theorem O.1 that there is a one-to-one map between RDEs and PPEDs. We use RDE and PPED interchangeably in the following.

7.2. An anti-folk theorem

Our main result in this section is an anti-folk theorem in our setting, which shows that payoffs in PPEDs are bounded away from efficiency even as $\beta \to 1$. To state this theorem, we define a set Q_D, which, in the spirit of FL, will capture the limit set of the PPED payoff sets $E_D(\beta)$ as $\beta \to 1$. To define Q_D, we first define the score analogously to FL.

Definition 8. For a strategy profile α and a direction $\lambda \in \mathbb{R}^2$, define $k_D(\alpha, \lambda, \beta)$ as the value of the program

$$\begin{aligned} &\sup_{v, w(y), D} \quad \lambda \cdot v \\ \text{s.t.} \quad & v = (1-\beta) g(\alpha) + \beta \mathbb{E}_{\pi(\alpha)[y]} \tilde{w}(y) \\ & v_i = (1-\beta) g_i(a_i, \alpha_{-i}) + \beta \mathbb{E}_{\pi(a_i, \alpha_{-i})[y]} \tilde{w}_i(y) \quad \forall a_i \in \operatorname{supp} \alpha_i \\ & v_i \geq (1-\beta) g_i(a_i, \alpha_{-i}) + \beta \mathbb{E}_{\pi(a_i, \alpha_{-i})[y]} \tilde{w}_i(y) \quad \forall a_i \in A_i \\ & \lambda \cdot v \geq \lambda \cdot w(y) \qquad \forall y \in Y \\ & D \in \mathcal{D}^2 \text{ is such that } (w, \tilde{w}, D) \text{ is consistent.} \end{aligned} \tag{4}$$

Define $k_D^*(\lambda, \beta) \equiv \sup_\alpha k_D(\alpha, \lambda, \beta)$.

It is easy to see that $k_D(\alpha, \lambda, \beta)$ is independent of β by the same scaling argument as in FL. We thus write $k_D^*(\lambda)$ instead of $k_D^*(\lambda, \beta)$. We define Q_D as

$$Q_D \equiv \bigcap_{\lambda \in \mathbb{R}^2} \left\{ v \in \mathbb{R}^2 : \lambda \cdot v \leq k_D^*(\lambda) \right\}. \tag{5}$$

The following result formalizes the aforementioned connection between Q_D and the limit set of PPED payoffs.

Theorem 4. *For all β, $E_D(\beta) \subseteq Q_D$. Moreover, if Q_D has full dimension (in $\mathbb{R}^2$), then $E_D(\beta) \to Q_D$ as $\beta \to 1$.*

There are numerous "folk theorem" results in the literature, in which equilibria of supergames can attain any payoff in the set V^* of feasible and individually rational payoff for sufficiently high β. In our context,

$$V^* \equiv \{v \in \text{co}(\{g(a)\}_a),\ v \geq \underline{v} \text{ elementwise}\}$$

where $\underline{v}$ is the minmax payoff of the stage game. As we will show next, however, in our case Q_D can never include all of V^*, which together with the inclusion $E_D(\beta) \subseteq Q_D$ implies an *anti*-folk theorem.

Theorem 5 *(Anti-folk theorem). Assume $\pi(a)$ has full support. Then, for every point $v \in Q_D$ that cannot be supported by a stage game Nash equilibrium, there is a point $v' \in V^*$ which strictly Pareto dominates v.*

This anti-folk theorem illustrates that our PPED solution concept leads to fundamentally different limit sets Q_D than the limit PPE set. Even in situations which satisfy the PPE folk theorem assumptions in FLM, the PPED limit set Q_D is bounded away from the Pareto frontier.

They key intuition for this result is as follows. In order to enforce an action that is very close to the Pareto frontier, it is necessary that punishments are almost parallel to the Pareto frontier (and hence barely waste any resources on Pareto inefficient continuation values). In a PPED, however, continuation values must be aligned and lie on a *positively* sloped line, preventing them to line up with the *negatively* sloped Pareto frontier. We prove our anti-folk theorem by first developing a recursive characterization of $E_D(\beta)$ following Abreu et al. (1990) and then using the properties of the score characterization of Q_D (i.e., the definition in (5)) from FL. We relegate the details of this proof to Online Appendix C.

7.3. PPED and strongly symmetric equilibria

In Definition 8, incentives are given by perceived payoffs $\tilde{w}(y)$, which, as was shown in Theorem 1, lie on a line with positive slope. This suggests a natural connection of our PPED concept with standard PPEs where continuation values are on a positively sloped line—a concept we will refer to as *linear PPE* or *PPEL*. We formally define PPEL as follows.

Definition 9 *(Linear PPE).* A PPE is said to be linear, denoted PPEL, if after each public history y^t the continuation values $\{w(y^t, y)\}_y$ lie on a positively sloped line. $E_L(\beta)$ is the set of payoffs of PPELs, given a discount factor β.

Note that when the stage game is symmetric, linear PPEs are a generalization of strongly symmetric equilibria (SSEs). In a symmetric game, a strongly symmetric equilibrium requires both players play the same strategies after every history, which implies that payoffs after each history are identical across players. Payoffs need not be identical across players in a linear PPE, but it must be the case that if one player "prefers" a history, so must his opponent, and in this sense it generalizes the concept of SSEs. Similar to our treatment of PPED in the previous subsection, we can define a linear score k_L and the limit PPEL set Q_L (see C.5).

With these definitions in mind, we state our main result in this section: under weak conditions on the stage game, PPELs and PPEDs are payoff-equivalent.

Theorem 6. *For any β, $E_L(\beta) \subseteq E_D(\beta)$, and $Q_L \subseteq Q_D$. Moreover, for any stage game Nash equilibrium payoff profile $v_{NE} \in \mathbb{R}^2$, $Q_D \cap \{v \geq v_{NE}\} \subseteq Q_L$. Thus, if the minmax payoff pair can be supported as a stage game Nash equilibrium, then $Q_D = Q_L$.*

Theorem 6 shows that conditional on lying to the top right of a stage game Nash equilibrium payoff pair, Q_L and Q_D coincide. In that sense, our PPED equilibrium concept, which unlike both PPEL and SSEs has a microfoundation that we have developed in this paper, is a generalization of strongly symmetric equilibria.

7.4. Examples of Q_D

In this section, we present examples of Q_D sets.[29] We will restrict our attention to prisoner's dilemmas. The main observations from this section are that (i) Q_D can consist of more than simply the Nash equilibrium payoffs, (ii) Q_D is bounded away from the efficient frontier of V^*, and (iii) changes in the signal structure that may expand the set of equilibrium payoffs under standard imperfect public monitoring may actually *decrease* the set of perceived payoffs in our setting. The second observation corresponds to Theorem 5, and the final one mirrors the ones presented in Section 4.

Consider two separate cases of the prisoner's dilemma with three public signals. The payoff matrix is the same in both games, but the signal structure differs. The tables below list payoffs and $\pi(\alpha)$.

	C	D
C	$1, 1$	$-1, 1.1$
D	$1.1, -1$	$0, 0$

$S3$	y'	y''
(C, C)	$1/2$	$1/2$
(D, C)	$1/4$	$3/4$
(C, D)	$1/4$	$3/4$
(D, D)	0	1

$S4$	y'	y''
(C, C)	$1/2$	$1/2$
(D, C)	$3/4$	$1/4$
(C, D)	$1/4$	$3/4$
(D, D)	0	1

The relation between S3 and S4 mirrors that between S1 and S2. In S3, y'' is more likely as long as either player defects. In S4, y' and y'' are equally likely if both players cooperate, but if player 1 defects, then y' is more likely, while if Player 2 defects, y'' is more likely.

Consider Game S4 first. Without ambiguity aversion, it can easily be checked that the score in the direction $\lambda = (1, 1)$ is 2, as the strategy profile (C, C) can be supported by the continuation payoffs $w(y') = (0.8, 1.2)$ and $w(y'') = (1.2, 0.8)$ for $\beta = 1/2$. In fact, despite that the full rank conditions are not satisfied, it turns out that the entire set of feasible, individually rational payoffs can be supported by a PPE. Now consider the same signal structure, but consider PPEDs. Suppose the w were in a *negatively* sloped line at the optimum in the linear program. In this case, it is easy to see that there would be no incentives for at least one player: either both $\tilde{w}$ would collapse to a point, or the $\tilde{w}$ would lie on a horizontal (or vertical) line. Thus, the only action profile that could be supported with w in a negatively sloped line is (D, D). Now suppose instead that the w were on a *positively* sloped line, and suppose without loss that $w(y') \geq w(y'')$. Then, we can choose the distortions D such that $\tilde{w} = w$, so we can just work with these continuation payoffs. Suppose we are trying to support a strategy where player i plays C with probability $\alpha_i > 0$. Then, we must have

$$\alpha_2 \cdot 1 + (1 - \alpha_2) \cdot (-1) + \alpha_2 \left(\frac{1}{2} w_1(y') + \frac{1}{2} w_1(y'') \right) + (1 - \alpha_2) \left(\frac{1}{4} w_1(y') + \frac{3}{4} w_1(y'') \right)$$
$$= \alpha_2 \cdot 1.1 + (1 - \alpha_2) \cdot 0 + \alpha_2 \left(\frac{3}{4} w_1(y') + \frac{1}{4} w_1(y'') \right) + (1 - \alpha_2) w_1(y''),$$

[29] We compute these sets by using Program O.16, which is easily implementable. Together with the observation that the prisoner's dilemma has a minmax Nash, Theorem 6 shows that Q_D coincides with Q_L.

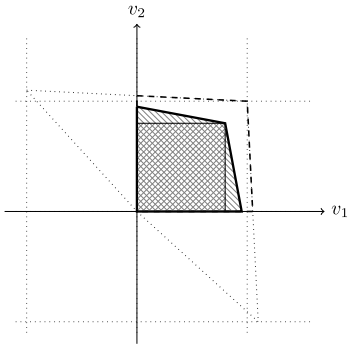

Fig. 7. The limit PPE set Q (single-hatched) and the limit PPED set Q_D (double-hatched) for a prisoner's dilemma with the aligned signal structure S3 given in the text.

which means that $0.9\alpha_2 - \left(\frac{3}{4} + \frac{\alpha_2}{2}\right)\left(w_1(y') - w_1(y'')\right) = 1$, and the equal sign is replaced by $\geq$ if $\alpha_1 = 1$. But, $w_1(y') - w_1(y'') \geq 0$ by assumption, and $0.9\alpha_2 \leq 0.9$, so this equation can never hold.[30] Thus, the score in the direction $\lambda = (1, 1)$ is 0, and $\{(0, 0)\}$ is the set of perceived PPED payoffs.

Now suppose that the signal structure is "aligned" as in Game S3. It can be checked that under standard preferences, the maximum score in the $\lambda = (1, 1)$ direction is 1.6, given by the profile (C, C) and the continuation payoffs $w(y') = (0.8, 0.8)$ and $w(y'') = (0.4, 0.4)$ for $\beta = 1/2$. That is, the score *decreases* relative to Game S3, as does the set of PPE payoff that can be supported as $\beta \to 1$. Next consider PPEDs. By the same argument as before, continuation payoffs (w) that lie on a negatively sloped line give no incentives $(\tilde{w})$, so we must restrict ourselves to choosing continuation payoffs on a positively sloped line if we want any hope of supporting anything more than the stage Nash equilibrium. But, note that the score then coincides with that in the case of regular PPEs (since we can choose the distortion strategies to set $w = \tilde{w}$) meaning that it *increases* relative to Game S4. Fig. 7 plots the limit PPE set Q, the limit PPED set Q_D, and V^* for Game S4: the dotted area is the convex hull of the payoff set, the dashed area is V^*, the single-hatched area is Q, and the criss-crossed area (also included in the single-hatched area) is Q_D. Note that Q_D is also bounded away from efficiency.

8. Conclusion

In this paper, we have proposed an explicit model of signal distortion that captures uncertainty over the timing and the technology of distortion. We show that if players are ambiguity averse over this timing and the distortion technology available to their opponents when they do so—or alternatively, choose actions that are robust to this uncertainty—then the "perceived incentives" (i.e., the perceived continuation values) lie on a positively sloped line. The intuition is relatively simple: given that a player will only change the signal to something that will benefit himself, the other player can ensure that she is made no worse off by this possibility of distortion if continuation payoffs are perfectly aligned. We also show that the intuition is robust to other natural models of signal distortion: these models both feature ambiguity aversion over the opponent's distortion technology but differ in the beliefs over the timing of these distortions.

[30] Obviously, if we assumed that $w(y'') \geq w(y')$ instead, then we would look at the decision rule for player 2.

We also present some examples that highlight differences between our setup and standard ones without this fear of signal distortion. First, signal structures that make deviators less distinguishable may in fact be beneficial for cooperation. Secondly, the possibility of distorting the signal can itself sustain cooperation when it is not possible in a standard setting. We use the prisoner's dilemma to illustrate this intuition, and we leave pursuing a generalization of these observations for future work.

Given that the proposed method of signal distortion is novel to game theory, we see many other possibilities for future research as well. First, our emphasis in this paper has been on "perceived payoffs." It would be interesting to extend our results to describe the behavior of actual payoffs under the true probabilities, through the eyes of an agent without ambiguity aversion. Second, we endogenize the continuation payoffs in our model through a repeated game, but there are other appealing methods to endogenize these payoffs. For instance, the $w(y)$ in the one-period model can be interpreted as payoffs set by a principal who wishes to motivate a team of agents who have a fear of signal distortion as presented in this paper. We view this line of research, connecting back to contract design, as especially promising. Finally, and more generally, we think that the concepts of distortion equilibria we introduced in this paper could be derived from a single generalization of extensive form games with imperfect information to the case where some agents are ambiguity averse about future draws by Nature. We speculate that even in such general settings, some kind of "alignment" may appear as agents naturally prefer branches of the game tree at which payoffs are aligned when they are ambiguity averse. We leave this extension of ambiguity aversion to general extensive form games for future research.

Acknowledgments

Thanks to Drew Fudenberg for encouraging us to undertake this project and for his guidance along the way. We are also grateful for helpful discussions with Gabriel Carroll, Glenn Ellison, Jonathan Libgober, David Miller, and Juuso Toikka—as well as for comments from the associate editor and three anonymous referees. All errors are our own. Bhattacharya and Manuelli acknowledge financial support from the NSF Graduate Research Fellowship under Grant No. 1122374. Straub acknowledges financial support from the ERP Fellowship ("European Recovery Fund Fellowship", sponsored by "Studienstiftung des Deutschen Volkes").

Appendix A. Extension to $N > 2$ players

In this appendix, we first extend our baseline model of Section 2 to $N > 2$ agents, and prove the analogue of Theorem 1 for a case with $N = 3$ players, which involves a direct extension of arguments presented in Section 3.1. In the second subsection, we show how to extend the model in Section 5.2 to arbitrarily many players. We then show that a many-agent version of our main linearity result carries over to this model.

A.1. Model of Section 2

In this section, we define a distortion equilibrium for $N > 2$ players. We then prove the analogue of Theorem 1 for a case with $N = 3$ players, which involves a direct extension of arguments presented in Section 3.1.

A distortion equilibrium as defined in Section 2.2 involves strategies for the first stage game along with distortion strategies for how players are willing to alter the signal, with these distortion strategies being disciplined by consistency in Definition 1. Implicit in this formulation is some notion of what sort of distortions players fear from their opponents. In the case of $N = 2$ players, there is not much choice in how to model this set: we simply assume that each player i fears that their opponents will distort after them and thus choose the signal y may yield any distortion in $D_{-i}(y)$.[31] However, for $N \geq 3$ players, this modeling choice is not obvious. Each player may worry that *exactly* one player will be able to distort after him, but he may be unsure which player it is. Each player may worry that all other players will be able to distort after him but may be uncertain about the sequence. Players may worry that all remaining $N - 1$ players will be able to distort the signal jointly.

We will add this modeling choice to the definition of the game. Note that in the $N = 2$-player setup of Section 2, the game could be summarized by the triple (G, w, π) of first-stage payoffs $G : A \to \mathbb{R}^N$, continuation payoffs $w : Y \to \mathbb{R}^N$, and a signal distribution $\pi : A \to \Delta Y$. In the general N-player setup, the primitives of the game will be a quadruple (G, w, π, Γ). The new component $\Gamma : \mathcal{D}^N \to \mathcal{D}^N$ takes in the players' distortion strategies and outputs the set of future distortions each player fears. For instance, in the baseline model in Section 2 with $N = 2$, we would have $\Gamma(D_1, D_2) = (D_2, D_1)$. The following are natural examples of Γ for $N > 2$.

I. Suppose each player fears that exactly one player will be able to distort after him but is unsure who. Then, $\Gamma_i(D)(y) = \bigcup_{j \neq i} D_j(y)$. This is because any μ that *some* player will accept in lieu of y is a possible future distortion from y that player i fears.
II. Suppose $N = 3$ and each player fears that exactly two players will distort and is sure about the order. For concreteness, say that player 1 believes with certainty that player 2 will get the chance to distort after him, and player 3 will distort after player 2. Then, the set of distortions that player 1 will fear is that of all possible *paths* of distortions. That is, $\Gamma_1(D)(y)$ is the set of all $\mu \in \Delta Y$ such that there exists $\mu_2 \in D_2(y)$ and $\mu_3(y') \in D_3(y')$ for all y' such that μ is the product of the row vector that represents μ_2 and the stacked row vectors that represent $\mu_3(y')$ for all y'.[32] Define $(D_2 * D_3)(y)$ as the value of $\Gamma_1(D)(y)$ in this case. (To understand this set better, note that $D_2(y)$ and $D_3(y)$ are both obviously subsets of $(D_2 * D_3)(y)$.)
III. If $N = 3$ and each player fears that exactly two players will distort but is unsure about the order, then $\Gamma_1(D)(y) = (D_2 * D_3)(y) \cup (D_3 * D_2)(y)$.
IV. If each player fears that the remaining $N - 1$ players will all jointly distort a signal after he does, then $\Gamma_i(D)(y) = \bigcap_{j \neq i} D_j(y)$. This is because any μ that player i fears must be preferred to y by *all* of his opponents.

Of course, it is simple to consider models where players fear multiple such possibilities by taking unions (just as for moving from Case II to Case III). However, some cases are clearly less restrictive than others. If player i fears both a single opponent distorting the signal and the possibility of all opponents jointly deciding to distort the signal, we would set $\Gamma_i(D)(y) =$

[31] As we note in Section 3.3, there is *some* degree of flexibility in how to model the distortion phase, since one may imagine a model in which players can distort multiple times. However, as long as each player fears a potential future distortion from his opponent at all times, nothing of import changes.

[32] We can think of this as a two-step non-homogenous Markov process on signals y, where player 2's distortion μ_2 gives the first transition probability vector and the a selection of distortions $\mu_3(y')$ for player 3 gives the second transition matrix.

$\bigcup_{j\neq i} D_j(y) \cup \bigcap_{j\neq i} D_j(y) = \bigcup_{j\neq i} D_j(y)$. This is expected, since collusion allows for fewer distortions than under individual distortions. Finally, note that the relevant Γ for each of the above cases can be derived by starting from an extensive form game like in Fig. 1 and adding ambiguity aversion.

With this formulation in mind, we modify the definitions from Section 2.2.

Definition 10 *(Consistency).* A triple $(w, \tilde{w}, D)$ is said to be consistent with respect to Γ if

$$\tilde{w}_i(y) = \min_{\mu\in\Gamma_i(D)(y)} \mathbb{E}_{\mu[y']} w_i(y) \text{ and}$$
$$D_i(y) = \{\mu \in \Delta(Y) : \mathbb{E}_{\mu[y']}[\tilde{w}_i(y')] \geq \tilde{w}_i(y)\}.$$

We can define distortion equilibrium analogously to before as well.

Definition 11 *(Distortion equilibrium).* A strategy profile (α, D) is a distortion equilibrium in the game (G, w, π, Γ) if

(i) the triple $(w, \tilde{w}, D)$ is consistent with respect to Γ as in Definition 10; and
(ii) for each player, α_i is optimal given α_{-i} and $\tilde{w}_i$, meaning for all $a_i \in \operatorname{supp}\alpha_i$,

$$a_i \in \arg\max_{a_i'\in A_i} (1-\beta) g_i(a_i', \alpha_{-i}) + \beta \cdot \mathbb{E}_{\pi_y(a_i',\alpha_{-i})[y]} \tilde{w}_i(y).$$

We now restrict to the case where $N = 3$ and explore the analogue of Theorem 1 for the possibilities for Γ presented above. We include the (somewhat repetitive) proofs to highlight that arguments very similar to ones in Section 3.1 apply. We leave an analysis of $N > 3$ for future work.

Lemma 3. *Suppose $\tilde{w}(y)$ and D satisfy consistency and Γ satisfies any of Cases I–IV above. Then, the $\tilde{w}(y)$ are strongly Pareto-ranked in that $\tilde{w}_i(y') \geq \tilde{w}_i(y'')$ if and only if $\tilde{w}_j(y') \geq \tilde{w}_j(y'')$ for any players i and j.*

Proof. Note that for any $\tilde{w}$ and pair of signals, at least two players must agree on the relative rankings of the signals. Thus, suppose without loss of generality that $\tilde{w}_1(y') \leq \tilde{w}_1(y'')$ and $\tilde{w}_2(y') \leq \tilde{w}_2(y'')$; Then, we have that $D_1(y'') \subseteq D_1(y')$ and $D_2(y'') \subseteq D_2(y')$. Then, it is easy to check that in each of Cases I–IV, we have $\Gamma_3(D)(y'') \subseteq \Gamma_3(D)(y')$. This implies that

$$\tilde{w}_3(y') = \min_{\mu\in\Gamma_3(D)(y')} \mathbb{E}_{\mu[y]} w_3(y) \leq \min_{\mu\in\Gamma_3(D)(y'')} \mathbb{E}_{\mu[y]} w_3(y) = \tilde{w}_3(y''),$$

as needed. □

Lemma 4. *Suppose $\tilde{w}(y)$ and D satisfy consistency and, among the $\tilde{w}(y)$, we have a unique Pareto-best point $\tilde{w}_B$ such that $\tilde{w}_B \geq \tilde{w}(y)$ for all y and a unique Pareto-worst point $\tilde{w}_W$ such that $\tilde{w}_W \leq \tilde{w}(y)$ for all y. Suppose further that Γ satisfies any of Cases I–III above. Then, the $\tilde{w}(y)$ can be expressed as $\tilde{w}(y) = c + t(y) \cdot d$, where $c \in \mathbb{R}^N$, $d \in \mathbb{R}^N$ with $d_i > 0$ for all i, and $t(y) \in \mathbb{R}$ for all y. That is, $\tilde{w}$ lie on a line with strictly positive slope.*

Proof. The proof proceeds almost exactly like the one to Lemma 2. If $\tilde{w}_{W,i} = \tilde{w}_{B,i}$ for any i, then Lemma 3 implies that it holds for *all* i.

Now suppose that $\tilde{w}_W < \tilde{w}_B$, and as in the proof of Lemma 2, let $Y_B \equiv \{y \in Y : \tilde{w}(y) = \tilde{w}_B\}$ and $Y_W \equiv \{y \in Y : \tilde{w}(y) = \tilde{w}_W\}$. Consider the space of $(\tilde{w}_1(y), \tilde{w}_2(y))$ and consider the line ℓ connecting $\tilde{w}_W$ to $\tilde{w}_B$ in this plane. Suppose that there exists $\hat{y}$ such that the project of $\tilde{w}(\hat{y})$ lies above ℓ in this plane. Note that for all i and $y_B \in Y_B$, we have that $D_i(y_B) = \Delta Y_B$, and for all i and $y_W \in Y_W$, we have $D_i(y_W) = \Delta Y$. In Cases I–III, this means that $\Gamma_i(D)(y_B) = \Delta Y_B$ for all $y_B \in Y_B$, and $\Gamma_i(D)(y_W) = \Delta Y$ for all $y_W \in Y_W$ too. Then, as in Footnote 12, we have that there exists y_B^* and y_W^* such that $w_2(y_B^*) = \tilde{w}_{B,2}$ and $w_2(y_W^*) = \tilde{w}_{W,2}$. Find the $\alpha \in [0, 1]$ such that $\tilde{w}_1(\hat{y}) = \alpha\tilde{w}_1(y_B^*) + (1-\alpha)\tilde{w}_1(y_W^*)$ and note that this means $\mu \equiv \alpha\delta_{y_B^*} + (1-\alpha)\delta_{y_W^*} \in D_1(\hat{y})$. For Cases I–III, this means that $\mu \in \Gamma_2(D)(\hat{y})$ as well. But, $\tilde{w}_2(\hat{y}) > \alpha\tilde{w}_2(y_B^*) + (1-\alpha)\tilde{w}_2(y_W^*) = \alpha w_2(y_B^*) + (1-\alpha)w_2(y_W^*)$, which is a contradiction to the definition of $\tilde{w}_2(\hat{y})$ as $\min_{\mu \in \Gamma_2(D)(\hat{y})} \mathbb{E}_{\mu[y]} w_2(y)$. Considering points to the right of ℓ leads to a similar contradiction. This shows that the projection of the points $\tilde{w}(y)$ onto the plane containing payoffs for players 1 and 2 lies on a strictly positive line. Since players 1 and 2 were arbitrary, □

Aggregating, we have the following analogue of Theorem 1.

Theorem 7. *If $N = 3$ and Γ is as in Cases I–IV, then consistency of the triple $(w, \tilde{w}, D)$ with respect to Γ requires that $\tilde{w}(y)$ are strongly Pareto-ranked. If Γ is as in Cases I–III, then they lie on a line of strictly positive slope.*

Note that we have not explored the most general conditions for Γ under which Lemmas 3 and 4 hold. For instance, Lemma 4 will hold for generalizations of Cases I–III for $N > 3$ as well. Instead, we have chosen to show the result for various natural assumptions on the distortion stage. For the case of $N = 3$, Cases I–IV (and combinations of thereof) likely exhaust all reasonable assumptions on this stage. The one case in which our linearity result may not hold (and incentives are instead simply Pareto-ranked) is Case IV. This case, however, is in some sense the one that embodies the lowest degree of fear of signal distortion since collusion places constraints on how opponents can decide to distort the signal. As long as each player fears *some* possibility of individual distortions by his opponents, we can show a linearity result.

A.2. Model of Section 5.2

We now turn to the model in Section 5.2. We will use the notation $T^{(i)}(w)$ to denote the perceived continuation values before Player i distorted, as a function of the continuation payoffs w after the distortion.

Generalizing the first stage game G, payoffs w, and signal distributions π to N players is straightforward, and analogous to the previous section. However, we now consider extensive forms similar to the one depicted in Fig. 6, just with an arbitrary sequence of distortion phases $(i_1, \ldots, i_J)$, where $J \geq 1$, and $i_j \in \{1, \ldots, N\}$ denotes the player distorting in distortion phase j. As in Section 5.2, our notion of distortion equilibrium is unchanged, with the exception that the perceived continuation values $\tilde{w}$ are now determined by a possibly long sequence of distortions,

$$\tilde{w} = \left(T^{(i_1)} \circ \cdots \circ T^{(i_J)}\right)(w). \tag{A.1}$$

We continue to call this equilibrium *distortion equilibrium*. In this environment, the following result holds.

Theorem 8. *Let $\tilde{w}$ be the perceived continuation values in a model with N players and J distortion phases by players $(i_1, \ldots, i_J)$.*

(a) *If there are $N-1$ distinct players that get to distort, $\#\{i_1, \ldots, i_J\} = N-1$, then all players' payoffs in $\tilde{w}$ are jointly Pareto ranked: if for some y, y' and i it holds that $\tilde{w}_i(y) < \tilde{w}_i(y')$, then $\tilde{w}_{i'}(y) \leq \tilde{w}_{i'}(y')$ for any other player i'.*
(b) *If all players get to distort, $\#\{i_1, \ldots, i_J\} = N$, then the points $\{\tilde{w}(y)\}$ lie on a line through $\mathbb{R}^N$ with a direction vector with positive entries.*

Proof. Before we begin our proof, we show four helpful properties of the map $T^{(i)}$. First, each image $\tilde{w} = T^{(i)}(w)$ is such that the vector $\tilde{w}_{-i}$ (consisting of all perceived payoffs other than the one for i) lies in the convex hull of the points in w_{-i}. This can be seen straight from the definition of consistent triple in Definition 4. In particular, if all points $w(y)$ lie on a line, then $\tilde{w}$ must lie on the exact same line.[33]

Second, for any w, and its image $\tilde{w} = T^{(i)}(w)$, $\{(\tilde{w}_i(y), \tilde{w}_{i'}(y))\}$ are Pareto-ranked for any other player i'. We will also say "the payoffs of players i and i' in $\tilde{w}$ are mutually Pareto-ranked" in that case. To see why, suppose $\tilde{w}_i(y) < \tilde{w}_i(y')$ for two signals y, y'. Then, because $\tilde{w}_i = w_i$, $D_i(y) \supseteq D_i(y')$, and hence by definition of $\tilde{w}$, $\tilde{w}_{i'}(y) \leq \tilde{w}_{i'}(y')$ for any other player i'. Now suppose that for some other Player $i' \neq i$ it holds that $\tilde{w}_{i'}(y) < \tilde{w}_{i'}(y')$. Clearly, if $\tilde{w}_i(y) > \tilde{w}_i(y')$, it must hold that $\tilde{w}_{i'}(y) \geq \tilde{w}_{i'}(y')$—a contradiction.

Third, we introduce the following notation: for any $\underline{x} \in \mathbb{R}$, define $B^{(i)}(w, \underline{x}) \equiv \text{co}(\{w(y)\}) \cap \{x \in \mathbb{R}^N : x_i \geq \underline{x}\}$. The $-i$ coordinates define the set of payoffs for players other than i that could result from a distortion of player i when faced with outside option $\underline{x}$. This lets us rewrite any image $\tilde{w} = T^{(i)}(w)$ as

$$\tilde{w}_{i'}(y) = \min_{x \in B^{(i)}(w, w_i(y))} x_{i'}, \tag{A.2}$$

an expression that holds for all players, including i.

Fourth, $T^{(i)} \circ T^{(i)} = T^{(i)}$. Let $\hat{w} = T^{(i)}(w)$ and $\tilde{w} \equiv T^{(i)}(\hat{w})$. To show that in fact $\tilde{w} = \hat{w}$, notice that $\tilde{w}_i = \hat{w}_i$ holds by construction. For any other i', we note that $\tilde{w}_{i'}(y) \leq \hat{w}_{i'}(y)$ since "no distortion" δ_y is always in the distortion set of agent i, $D_i(y)$. To prove the reverse inequality, we note that $B^{(i)}(\hat{w}, \underline{x}) \subseteq B^{(i)}(w, \underline{x})$ for any $\underline{x}$ by our first property, and therefore, using (A.2),

$$\tilde{w}_{i'}(y) = \min_{x \in B^{(i)}(\hat{w}, \hat{w}_i(y))} x_{i'} \geq \min_{x \in B^{(i)}(w, w_i(y))} x_{i'} = \hat{w}_{i'}(y).$$

Fifth, for any two distinct players i, i', and $\tilde{w} = \left(T^{(i)} \circ T^{(i')}\right)(w)$ it holds that $\{(\tilde{w}_i(y), \tilde{w}_{i'}(y))\}$ lie on a line. We will also say "the payoffs of players i and i' in $\tilde{w}$ are mutually aligned". This property can be proved analogous to Theorem 3 after relabeling i, i'.

Having established these five properties, we now turn to the actual proof of the theorem. Let N_J be the subset of agents that get to distort the signal, and let $n \equiv \#N_J \leq N$ be the number of distinct players that get to distort the signal. Define $\tilde{w}$ as in (A.1). Due to properties 1 and 5, all n players' payoffs in $\tilde{w}$ must be pairwise mutually aligned. But this means their payoffs must also be jointly linear, in the sense that the points $\{(w_i(y))_{i \in N_J}\}_{y \in Y}$ lie on a line in $\mathbb{R}^n$ with strictly positive direction vector.

[33] In fact one can show that in that case $T^{(i)}(w) = w$ but we do not require this fact for the proof at hand.

If $n = N$, this already proves statement (b) above. If $n = N - 1$, a single player does not get to distort the signal. For simplicity, label that player "Player 1", and label the first player to distort "Player 2", i.e. $i_1 = 2$. Then by property 2, the payoffs of players 1 and 2 in $\tilde{w}$ are mutually Pareto-ranked. It is easy to see that, together with the joint linearity among players $2, \ldots, N$, this implies that all N agents' payoffs in $\tilde{w}$ are jointly Pareto-ranked. □

Appendix B. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jet.2018.01.004.

References

Abreu, D., Pearce, D., Stacchetti, E., 1986. Optimal cartel equilibrium with imperfect monitoring. J. Econ. Theory 39 (1), 251–269.
Abreu, D., Pearce, D., Stacchetti, E., 1990. Toward a theory of discounted repeated games with imperfect monitoring. Econometrica 58 (5), 1041–1063.
Bagwell, K., Staiger, R.W., 1990. A theory of managed trade. Am. Econ. Rev. 80 (4), 779–795.
Bodoh-Creed, A.L., 2012. Ambiguous beliefs and mechanism design. Games Econ. Behav. 75 (2), 518–537.
Bose, S., Ozdenoren, E., Pape, A., 2006. Optimal auctions with ambiguity. Theor. Econ. 1 (4), 411–438.
Bose, S., Renou, L., 2014. Mechanism design with ambiguous communication devices. Econometrica 82 (5), 1853–1872.
Carroll, G., 2015. Robustness and linear contracts. Am. Econ. Rev. 105 (2), 536–563.
Chassang, S., 2013. Calibrated incentive contracts. Econometrica 81 (5), 1935–1971.
Chung, K.-S., Ely, J.C., 2007. Foundations for dominant strategy mechanisms. Rev. Econ. Stud. 74 (2), 447–476.
Di Tillio, A., Kos, N., Messner, M., 2017. The design of ambiguous mechanisms. Rev. Econ. Stud. 84 (1), 237–276.
Dow, J., da Costa Werlang, S.R., 1994. Nash equilibrium under Knightian uncertainty: breaking down backward induction. J. Econ. Theory 64 (2), 302–324.
Edmans, A., Gabaix, X., 2011. Tractability in incentive contracting. Rev. Financ. Stud. 24 (9), 2865–2894.
Fudenberg, D., Levine, D.K., 1994. Efficiency and observability with long-run and short-run players. J. Econ. Theory 62 (1), 103–135.
Fudenberg, D., Levine, D., Maskin, E., 1994. The folk theorem with imperfect public information. Econometrica 62 (5), 997–1039.
Fudenberg, D., Yamamoto, Y., 2010. Repeated games where the payoffs and monitoring structure are unknown. Econometrica 78 (5), 1673–1710.
Fudenberg, D., Yamamoto, Y., 2011. Learning from private information in noisy repeated games. J. Econ. Theory 146 (5), 1733–1769.
Gilboa, I., Schmeidler, D., 1989. Maxmin expected utility with non-unique prior. J. Math. Econ. 18 (2), 141–153.
Green, E.J., Porter, R.H., 1984. Noncooperative collusion under imperfect price information. Econometrica 52 (1), 87–100.
Holmström, B., Milgrom, P., 1987. Aggregation and linearity in the provision of intertemporal incentives. Econometrica 55 (2), 303–328.
Kandori, M., 1992. The use of information in repeated games with imperfect monitoring. Rev. Econ. Stud. 59 (3), 581–593.
Klibanoff, P., 1996. Uncertainty, Decision, and Normal Form Games. Working paper. Northwestern University.
Levin, J., 2003. Relational incentive contracts. Am. Econ. Rev. 93 (3), 835–857.
Lo, K.C., 1996. Equilibrium in beliefs under uncertainty. J. Econ. Theory 71 (2), 443–484.
Lo, K.C., 1999. Extensive form games with uncertainty averse players. Games Econ. Behav. 28 (2), 256–270.
Lopomo, G., Rigotti, L., Shannon, C., 2010. Uncertainty in Mechanism Design. Working paper. Duke University.
Lopomo, G., Rigotti, L., Shannon, C., 2011. Knightian uncertainty and moral hazard. J. Econ. Theory 146 (3), 1148–1172.
Marshall, R.C., Marx, L.M., 2012. The Economics of Collusion: Cartels and Bidding Rings. MIT Press, Cambridge, MA.
Ozdenoren, E., Peck, J., 2008. Ambiguity aversion, games against nature, and dynamic consistency. J. Econ. Theory 62 (1), 106–115.

Radner, R., 1986. Repeated partnership games with imperfect monitoring and no discounting. Rev. Econ. Stud. 53 (1), 43–57.
Sannikov, Y., Skrzypacz, A., 2010. The role of information in repeated games with frequent actions. Econometrica 78 (3), 847–882.
Wolitzky, A., 2016. Mechanism design with maxmin agents: theory and an application to bilateral trade. Theor. Econ. 11 (3), 971–1004.