Games and Economic Behavior 117 (2019) 15–39

Contents lists available at ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

Frustration, aggression, and anger in leader-follower games ☆

Pierpaolo Battigalli [a], Martin Dufwenberg [b,c,d], Alec Smith [e,*]

[a] Bocconi University and IGIER, Italy
[b] University of Arizona, USA
[c] University of Gothenburg, Sweden
[d] CESifo, Germany
[e] Virginia Tech, Department of Economics, USA

ARTICLE INFO

Article history:
Received 26 October 2017
Available online 20 June 2019

JEL classification:
C72
D01
D91

Keywords:
Frustration
Anger
Blame
Belief-dependent preferences
Psychological games
Threats

ABSTRACT

Frustration, anger, and blame have important consequences for economic and social behavior, concerning for example monopoly pricing, contracting, bargaining, violence, and politics. Drawing on insights from psychology, we develop a formal approach to exploring how frustration and anger, via blame and aggression, shape interaction and outcomes in a class of two-stage games.

1. Introduction

Anger can shape economic outcomes. Consider three cases:

Case 1: In 2015 Turing Pharmaceuticals raised the price of Daraprim, a therapeutic drug, from $12 to $750 per dose. The company was subsequently accused of price gouging. Should Turing have considered the consequences of customer anger before determining the new price for the drug?

Case 2: When local football teams favored to win instead lose, the police get more reports of husbands assaulting wives (Card and Dahl, 2011). Do unexpected losses spur vented frustration?

Case 3: Following the sovereign debt crises that began in 2009, some EU countries embarked on austerity programs. Was it because citizens lost benefits that some cities experienced riots?

☆ This paper modifies and extends Smith (2009). Pierpaolo Battigalli gratefully acknowledges financial support from ERC advanced grant 324219. We thank Chiara Aina, Federico Bobbio, Roberto Corrao, Carlo Cusumano, Giacomo Lanzani, Paolo Leonetti, Paola Moscariello, and Marco Stenborg Petterson for excellent research assistance, and Doug Bernheim, Steve Brams, Leda Cosmides, Vince Crawford, Nicodemo De Vito, Uri Gneezy, Pierfrancesco Guarino, Michele Griessmair, Heike Hennig-Schmidt, Botond Kőszegi, Alex Imas, Joshua Miller, David Rietzke, Julio Rotemberg, Emanuel Vespa, several reviewers, and many seminar and conference audiences for helpful comments.

* Corresponding author.
E-mail addresses: pierpaolo.battigalli@unibocconi.it (P. Battigalli), martind@eller.arizona.edu (M. Dufwenberg), alecsmith@vt.edu (A. Smith).

https://doi.org/10.1016/j.geb.2019.06.001
0899-8256/

Pricing, domestic violence, political landscapes: these are important themes. We propose that others (involving—say—recessions, contracting, arbitration, terrorism, road rage, or support for populists) could plausibly be imagined. However, to assess the impact of anger on social and economic interactions, one needs a theory that predicts outcomes based on the decision-making of anger-prone individuals and that accounts for strategic considerations. We develop such a theory and focus on **leader-follower (LF-) games** with two players, one of whom moves first, while the other observes and reacts.

Insights from psychology about the triggers and repercussions of anger are evocative. The behavioral consequences of emotions are called "action tendencies," and the action tendency of anger is aggression and urge to retaliate. Angry players may be willing to forgo material gains to punish others, or be predisposed to aggression when this serves as a credible threat. But while insights of this nature can be gleaned from psychologists' writings, their analysis usually stops with the individual rather than going on to assess overall economic implications. We take the basic insights about anger that psychology has produced as input and inspiration.[1]

We study the strategic interaction of decision makers who become angry when they are frustrated.[2] Frustration occurs when someone is unexpectedly denied something he or she cares about. We assume that people are frustrated when they get less material rewards than expected.[3] They then become hostile towards whomever they blame. Because a player's frustration depends on his beliefs about others' choices, and the blame a player attributes to another may depend on his beliefs about others' choices or beliefs, all our models find their intellectual home in the framework of psychological game theory; see Geanakoplos et al. (1989) and Battigalli and Dufwenberg (2009) (B&D).

Initially expected material payoffs are the reference point to which outcomes are compared to generate frustration. This modeling choice mirrors several other behavioral models, including Kőszegi and Rabin's (2006, 2007) model of reference-dependent preferences, guilt aversion (Dufwenberg, 2002, Battigalli and Dufwenberg, 2007), and earlier models of disappointment aversion (Bell, 1985, Loomes and Sugden, 1986). While this approach may not incorporate every aspect of frustration discussed in the psychology literature, it is consistent with a range of phenomena, allowing for frustration and anger to be belief-dependent and capturing the stylized fact that costly punishment can violate consequentialism (*e.g.*, Falk et al., 2003, 2008).

There are many ways to model blame.[4] We present three approaches that result in distinct utilities. Players motivated by simple anger (SA) become generally hostile when frustrated. In contrast, those motivated by anger from blaming behavior (ABB) or anger from blaming intentions (ABI) are more discriminating, asking who caused, or intended to cause, their frustration. SA captures the psychological phenomenon of *displaced aggression*, where an angry person takes out frustration on a blameless bystander.[5] However, for some authors, blame or other-responsibility is a prerequisite for anger.[6] ABB and ABI pick that up.

Players have beliefs about both others' beliefs and actions as well as their own actions. Because frustration results from diminished expectations, first-movers are never frustrated at the root and maximize their expected material payoffs. We define and establish the existence of a notion of sequential equilibrium (SE) that adapts to our framework the one developed in B&D. In pure-strategy SE, frustration arises only off the equilibrium path, and furthermore, in generic perfect information game forms, there is always an SE with anger that is realization-equivalent to the material-payoff equilibrium, though anger may also result in additional equilibria. In LF-games, a follower with SA, ABB, or ABI always does at least as well, materially, as a player who is not anger-prone.[7] We also formally develop the notion of a **threat** in order to provide a partial characterization of SE with anger: the presence of threats allows anger-prone followers to obtain more than in the material-payoff equilibrium and give less to the leader, while their absence implies that equilibria with SA, ABB, or ABI are equivalent to the material payoff equilibrium.

Our models can encapsulate Cases 1 and 2 above (for Case 3, cf. Passarelli and Tabellini, 2017). Case 1 is captured by an ultimatum minigame (an example of an LF-game), where SA and ABB allow for a pure SE involving rejection of the greedy offer, and all our concepts allow for an SE where rejection occurs with positive probability. Case 2 is illustrated with SA in

[1] The psychology literature is huge. A source of inspiration is *International Handbook of Anger* (Potegal et al., 2010) offering a cross-disciplinary perspective reflecting "affective neuroscience, business administration, epidemiology, health science, linguistics, political science, psychology, psychophysiology, and sociology" (p. 3). The absence of "economics" in the list may indicate that our approach is original!

[2] A large body of work in psychology connects frustration, anger, and aggression, beginning with Dollard et al. (1939). See, for example, Averill (1982), Berkowitz (1989), and the (*op. cit.*) *Handbook*, especially the chapters by Lewis, Wranik and Scherer, and by Berkowitz.

[3] That frustration depends on expectations is well supported in the psychology literature, *e.g.*, Berkowitz (1978, p. 697): "Unlike deprivations, frustrations can only be surprising (to a greater or lesser extent). People who do not expect to reach their goals are not anticipating the pleasure these goals would bring. Their hopes are not dashed if they have no hopes." Our focus on material rewards is admittedly restrictive. See the Discussion in Section 5.

[4] See, *e.g.*, Alicke (2000), Battigalli and Dufwenberg (2007), and Halpern (2016, Chapter 6).

[5] See Marcus-Newhall et al. (2000).

[6] See, *e.g.*, the chapter by Wranik and Scherer in the (*op. cit.*) *Handbook*.

[7] Taking into account that evolutionary pressure is driven by material payoffs (*e.g.*, Buss, 2016), this result is consistent with the work of Sell et al. (2009), who argue that anger is the result of a process of natural selection for behaviors that resolve bargaining conflicts in favor of the anger-prone individual.

an example that we call "hammering-one's-thumb."[8] In contrast, incorporating notions of blame into our analysis, either of ABB and ABI eliminates such displaced aggression.

A small literature examines the role of anger in economic behavior. Selten (1978) discusses the implications of the frustration-aggression hypothesis of Dollard et al. (1939) for behavior in the chain store game, though he does not develop a model.[9] Other early work exploring the role of anger in solving commitment problems includes Hirshleifer (1987), Frank (1988), and Elster (1998). Most recent studies are empirical or experimental, indicative of hostile action occurring in economic situations, based on either observational data or experimental data.[10] A few studies present theories different from ours, including Rotemberg (2005, 2008, 2011), Brams (2011), Winter (2014), Winter et al. (2016), Akerlof (2016), and Passarelli and Tabellini (2017).

We present our three models in Section 2. Section 3 contains most of our analysis, defining SE and exploring results focused mainly on the class of LF-games. Section 4 compares our approach to others (those mentioned above, plus some models of distributional preferences and reciprocity). Section 5 contains a broad concluding discussion. Proofs are collected in the Appendix.

2. Three models

2.1. Preliminaries

Game form We restrict attention to finite two-stage game forms with observed actions, meaning that first-period choice profiles become public information at the beginning of the second period. The set of players is I. To ease notation, we assume that all players take actions simultaneously at each stage. Thus, nodes are histories h of action profiles $a^t = \left(a_i^t\right)_{i\in I}$, observed by all; $h = \varnothing$ is the empty history (the root), $h = (a^1)$ a history of length one, which may be terminal or not, and $h = (a^1, a^2)$ a history of length 2, which is terminal. H is the set of nonterminal histories and Z is the set of terminal histories (end nodes). The set of feasible actions of i after $h \in H$ is $A_i(h)$. This set is a singleton if i is not active given h. Thus, for all $h \in H$, $I(h) = \{i \in I : |A_i(h)| > 1\}$ is the set of active players following h. Given the observed actions property, if $I(h)$ is a singleton for each $h \in H$ then the game has **perfect information**. We omit parentheses whenever no confusion may arise. For example, we may write $h = a^1$ instead of $h = (a^1)$, and $h = (a_i^1, a_j^2)$ if i (resp. j) is the only first (resp. second) mover. Finally, we let $A(h) = \times_{i\in I} A_i(h)$ and $A_{-i}(h) = \times_{j\neq i} A_j(h)$. The material consequences of players' actions are determined by a profile of monetary payoff functions $\left(\pi_i : Z \to \mathbb{R}\right)_{i\in I}$. A *perfect information* game has **no relevant ties** if distinct terminal histories yield different payoffs for the player who is active at the longest common prefix.[11] For two-stage games, this means that different actions of the first mover lead to different material payoffs for the first mover, and different actions of a second mover lead to different material payoffs for this second mover. If the game contains chance moves, we augment the player set with a dummy player c (with $c \notin I$), who selects a feasible action at random. Thus, let $I_c = I \cup \{c\}$, and the sets of first and second movers may include c: $I(\varnothing), I(a^1) \subseteq I_c$. If the chance player is active at $h \in H$, its move is described by probability mass function $\sigma_c(\cdot|h) \in \Delta(A_c(h))$.

Personal histories To model how i determines the subjective value of feasible actions, we add to the commonly observed histories $h \in H$ also personal histories of the form (h, a_i), with $a_i \in A_i(h)$. In a game with perfect information, $(h, a_i) \in H \cup Z$. But if there are simultaneous moves at h, then (h, a_i) is not a history in the standard sense. As soon as i irreversibly chooses action a_i, he observes (h, a_i), and can determine the value of a_i using his beliefs conditional on this event (i knows in advance how he is going to update his beliefs conditional on what he observes). We denote by H_i the set of histories of i—standard and personal—and by $Z(h_i)$ the set of terminal successors of h_i.[12] The standard precedence relation $\prec$ for histories in $H \cup Z$ is extended to H_i in the obvious way: for all $h \in H$, $i \in I(h)$, $a_i \in A_i(h)$, and $a_{-i} \in A_{-i}(h)$ it holds that $h \prec (h, a_i)$ and, furthermore, $(h, a_i) \prec (h, (a_i, a_{-i}))$ if i is not the only active player at h. Note that $h_i \prec h_i'$ implies $Z(h_i') \subseteq Z(h_i)$, with strict inclusion if $h_i \in H$ and at least one player (possibly c) is active at h_i.

[8] The example is inspired by Frijda (1993), who says "Many experiences or responses of anger... are elicited by events that involve no blameworthy action" and suggests that simple frustrations such as "one's car refusing to start, finding one's bicycle has a flat tyre, rain on the fifth day of one's holiday after four previous days of rain... hitting one's head on the kitchen shelf, dropping a needle for the third time in a row" or "hammering one's thumb" may result in anger. He goes on to say that "the target of anger may be a person who has fallen ill on the day of one's party, or one who just happened to be present when a plan failed".

[9] The chain store stage game is strategically equivalent to the ultimatum minigame.

[10] See Anderson and Simester (2010) and Rotemberg (2005, 2011) on pricing; Card and Dahl (2011) and Munyo and Rossi (2013) on violence; Carpenter and Matthews (2012), Gurdal et al. (2014), Gneezy and Imas (2014), Persson (2018), van Leeuwen et al. (2018), Aina et al. (2018), and Dufwenberg et al. (2018a, 2018b) for experiments.

[11] See Battigalli (1997). It can be checked that the property is *generic* with respect to material payoff functions $\pi \in \mathbb{R}^{Z\times I}$: the closure of the set of π that do not satisfy it has Lebesgue measure 0 in $\mathbb{R}^{Z\times I}$.

[12] That is, $H_i = H \cup \{(h, a_i) : h \in H, i \in I(h), a_i \in A_i(h)\}$. The definition of $Z(h_i)$ is standard for $h_i \in H$; for $h_i = (h, a_i)$ we have $Z(h, a_i) = \bigcup_{a_{-i}\in A_{-i}(h)} Z(h, (a_i, a_{-i}))$.

Beliefs It is conceptually useful to distinguish three aspects of a player's beliefs: beliefs about co-players' actions, beliefs about co-players' beliefs, and the player's plan, which are beliefs about own actions. Beliefs are defined conditional on each $h_i \in H_i$. Abstractly denote by Δ_{-i} the space of co-players' beliefs (the formal definition is given below). Player i's beliefs can be described as conditional probability measures over paths and beliefs of others, *i.e.*, over $Z \times \Delta_{-i}$. Events, from i's point of view, are subsets: Events about behavior take form $Y \times \Delta_{-i}$, with $Y \subseteq Z$; events about beliefs take form $Z \times E_{\Delta_{-i}}$, with $E_{\Delta_{-i}} \subseteq \Delta_{-i}$. Here we provide an abridged and somewhat informal description of beliefs that is sufficient to understand the main text. The Appendix contains the formal analysis used in proofs.

First-order beliefs For each $h_i \in H_i$, player i holds beliefs $\alpha_i(\cdot|h_i) \in \Delta(Z(h_i))$ about the actions that will be taken in the continuation of the game. The system of beliefs $\alpha_i = (\alpha_i(\cdot|h_i))_{h_i \in H_i}$ must satisfy two properties. First, the rules of conditional probabilities hold whenever possible, that is, α_i satisfies the **chain rule** of conditional probabilities (Eq. (7), see the Appendix). Second, if at history $h \in H$ player i moves simultaneously with other players, what i believes about the *simultaneous* actions of co-player is *independent of his action*, that is, such marginal probabilities are the same conditional on h and on (h, a_i), for every $a_i \in A_i(h)$. We call this natural property **own-action independence** (see Eq. (8) in the Appendix).

To ease notation, for each $h \in H$ and each action profile $a = (a_i, a_{-i}) \in A(h)$, let $\alpha_i(a|h)$, $\alpha_{i,i}(a_i|h)$, and $\alpha_{i,-i}(a_{-i}|h)$ respectively denote the (marginal) conditional probabilities assigned by α_i to a, a_i, and a_{-i}. The chain rule and own-action independence imply

$$\alpha_i(a_i, a_{-i}|h) = \alpha_{i,i}(a_i|h)\alpha_{i,-i}(a_{-i}|h).$$

Thus, α_i is made of two parts, what i believes about own behavior *and* about the behavior of others. The array of probability measures $\alpha_{i,i} \in \times_{h \in H} \Delta(A_i(h))$ is—technically speaking—a behavior strategy, and we interpret it as i's **plan**. The reason is that the result of i's contingent planning is precisely a system of conditional beliefs about what action he would take at each history. If there is only one co-player, also $\alpha_{i,-i} \in \times_{h \in H} \Delta(A_{-i}(h))$ corresponds to a behavior strategy. With multiple co-players, $\alpha_{i,-i}$ corresponds instead to a "correlated behavior strategy." Whatever the case, $\alpha_{i,-i}$ gives i's conditional beliefs about others' behavior, and these beliefs may not coincide with the plans of others. We emphasize: a player's plan does not describe actual choices, actions on the path of play are the only actual choices. A conditional belief system α_i satisfying the chain rule and own-action independence is a **first-order** belief of i. Let Δ_i^1 denote the space of such beliefs. It can be checked that Δ_i^1 is a compact metric space, hence the same holds for $\Delta_{-i}^1 = \times_{j \neq i} \Delta_j^1$, the space of co-players' first-order beliefs profiles.

Second-order beliefs Players also hold beliefs about the beliefs of co-players. In the following analysis, the only co-players' beliefs affecting the values of actions are their first-order beliefs. Therefore, we limit our attention to **second-order** beliefs, *i.e.*, systems of conditional probability measures $\beta_i = (\beta_i(\cdot|h_i))_{h_i \in H_i} \in \times_{h_i \in H_i} \Delta\left(Z(h_i) \times \Delta_{-i}^1\right)$ that satisfy the chain rule (Eq. (9), in the Appendix) and the appropriate version of the own-action independence property for second-order beliefs: given any history $h \in H$, what i believes about the simultaneous action of co-players *and their first-order belief systems* conditional on h and conditional on (h, a_i) is the same for every $a_i \in A_i(h)$ (see Eq. (10) in the Appendix). The space of such second-order belief systems of i is denoted by Δ_i^2. It can be checked that the marginalization onto Z of any $\beta_i \in \Delta_i^2$ yields a system of first-order beliefs satisfying the chain rule and own action-independence, that is, an element α_i of Δ_i^1. This α_i is the first-order belief implicit in β_i. Whenever we write in a formula beliefs of different orders for i, we assume that α_i *is derived from* β_i, otherwise beliefs of different orders would not be mutually consistent. Also, we may omit the empty history, as in $\beta_i(E) = \beta_i(E|\varnothing)$ or $\alpha_i(a) = \alpha_i(a|\varnothing)$.

Conditional expectations Let ψ_i be any real-valued measurable function of variables that i does not know, *e.g.*, the terminal history or the co-players' first-order beliefs. Then i can compute the expected value of ψ_i conditional on any $h_i \in H_i$ by means of his belief system β_i, denoted $\mathbb{E}[\psi_i|h_i; \beta_i]$. If ψ_i depends only on actions, *i.e.*, on the path z, then $\mathbb{E}[\psi_i|h_i; \beta_i]$ is determined by the α_i derived from β_i, and we can write $\mathbb{E}[\psi_i|h_i; \alpha_i]$. In particular, α_i gives the conditional expected material payoffs:

$$\mathbb{E}[\pi_i|h; \alpha_i] = \sum_{z \in Z(h)} \alpha_i(z|h)\pi_i(z),$$

$$\mathbb{E}[\pi_i|(h, a_i); \alpha_i] = \sum_{z \in Z(h, a_i)} \alpha_i(z|h, a_i)\pi_i(z)$$

for all $h \in H$, $a_i \in A_i(h)$. $\mathbb{E}[\pi_i|h; \alpha_i]$ is what i expects to get conditional on h given α_i, which also specifies i's plan. $\mathbb{E}[\pi_i|(h, a_i); \alpha_i]$ is i's expected payoff of action a_i given h. If a_i is what i planned to choose at h, $\alpha_{i,i}(a_i|h) = 1$, and then $\mathbb{E}[\pi_i|h; \alpha_i] = \mathbb{E}[\pi_i|(h, a_i); \alpha_i]$. For initial beliefs, we omit $h = \varnothing$, writing $\mathbb{E}[\pi_i; \alpha_i]$ for i's initially expected material payoff.

Table 1 summarizes our framework.

Table 1
Elements of the two-stage game form.

Notation	Terminology
$i \in I$	players
$h \in H$	non-terminal, or partial histories
$I(h) \subseteq I$	set of active players at h
$t \in \{1,2\}$	stages, or periods
$A_i(h)$, $A(h)$, $A_{-i}(h)$	set of actions and action profiles at h
a_i^t	action of i in stage t
a^t (a_{-i}^t)	action profile (of others) in stage t
$z \in Z$	terminal histories
$Z(h)$	terminal successors of h
$\pi_i : Z \to \mathbb{R}$	monetary payoff function of $i \in I$
$\alpha_i, \alpha_{-i}, \alpha$	First-order beliefs and belief profiles
$\beta_i, \beta_{-i}, \beta$	Second-order beliefs and belief profiles

2.2. The frustration-aggression hypothesis and three ways to blame

Anger is triggered by frustration. While we focus on anger as a social phenomenon—frustrated players blame, become angry with, and care for the payoffs of *others*—our account of frustration refers to own payoffs only.[13] We define player i's **frustration** at history h, given his first-order belief system α_i as

$$\mathrm{F}_i(h;\alpha_i) = \left[\mathbb{E}[\pi_i;\alpha_i] - \max_{a_i \in A_i(h)} \mathbb{E}[\pi_i|(h,a_i);\alpha_i]\right]^+,$$

where $[x]^+ = \max\{x,0\}$. In words, frustration is given by the gap, if positive, between i's initially expected payoff and the currently best expected payoff he believes he can obtain. Diminished expectation—$\mathbb{E}[\pi_i|h;\alpha_i] < \mathbb{E}[\pi_i;\alpha_i]$—is only a necessary condition for frustration. For i to be frustrated it must also be the case that i cannot close the gap.

At the root, frustration must be zero; nothing happened, hopes can't be dashed:

Remark 1. For every player $i \in I$ and system of first-order beliefs $\alpha_i \in \Delta_i^1$, frustration must equal 0 at the initial history $h = \varnothing$, because the chain rule and own-action independence imply

$$\mathbb{E}\left[\pi_i;\alpha_i\right] = \sum_{a_i^1 \in A_i(\varnothing)} \alpha_{i,i}\left(a_i^1|\varnothing\right)\mathbb{E}\left[\pi_i|a_i^1;\alpha_i\right] \leq \max_{a_i^1 \in A_i(\varnothing)} \mathbb{E}\left[\pi_i|a_i^1;\alpha_i\right].$$

Frustration is possible at the end nodes, but can't influence subsequent choices as the game is over. One might allow the anticipation of frustration to be felt at end nodes to influence earlier decisions; however, the assumptions we make below rule this out. In our 2-stage setting, all behaviorally relevant frustration occurs in the second stage and is given by

$$\mathrm{F}_i(a^1;\alpha_i) = \left[\mathbb{E}[\pi_i;\alpha_i] - \max_{a_i^2 \in A_i(a^1)} \mathbb{E}[\pi_i|(a^1,a_i^2);\alpha_i]\right]^+.$$

Preferences at a given node depend on expected material payoffs and frustration. A frustrated player is motivated to hurt others, if this is not too costly (cf. Dollard et al., 1939; Averill, 1982; Berkowitz, 1989). We consider versions of this frustration-aggression hypothesis related to different cognitive appraisals of blame. In general, player i moving at h chooses action a_i to maximize the expected value of a belief-dependent "decision utility" of the form

$$u_i\left(h,a_i;\beta_i\right) = \mathbb{E}\left[\pi_i|\left(h,a_i\right);\alpha_i\right] - \theta_i \sum_{j \neq i} \mathrm{B}_{ij}\left(h;\beta_i\right)\mathbb{E}\left[\pi_j|\left(h,a_i\right);\alpha_i\right], \tag{1}$$

where α_i is derived from β_i, and $\theta_i \geq 0$ is a sensitivity parameter. Thus, $\mathrm{B}_{ij}\left(h;\beta_i\right) \geq 0$ measures how much of i's frustration is blamed on j, and the presence of $\mathbb{E}\left[\pi_j|\left(h,a_i\right);\alpha_i\right]$ in the formula translates this into a tendency to hurt j. We assume that

$$\mathrm{B}_{ij}(h;\beta_i) \leq \mathrm{F}_i(h;\alpha_i). \tag{2}$$

Because there is no frustration in the first stage, the following is true.

[13] In Section 5 (in hindsight of definitions to come) we discuss this approach in depth.

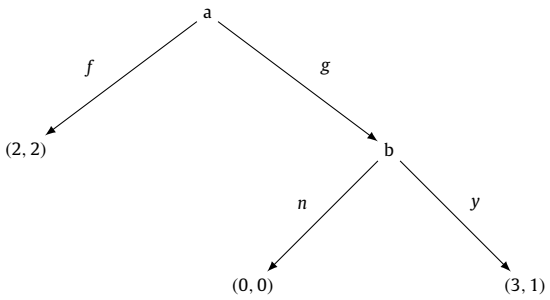

Fig. 1. Ultimatum minigame.

Remark 2. The decision utility of a first-mover coincides with expected material payoff: for all $i \in I(\varnothing)$, $a_i \in A_i(\varnothing)$, $\beta_i \in \Delta_i^2$, and $\alpha_i \in \Delta_i^1$ derived from β_i,

$$u_i(\varnothing, a_i; \beta_i) = \mathbb{E}[\pi_i | a_i; \alpha_i].$$

Simple anger (SA) Our most rudimentary hypothesis is that i's tendency to hurt others is proportional to i's frustration. SA is unmodulated by the cognitive appraisal (i.e., personal interpretation) of blame, so $\mathrm{B}_{ij}(h; \beta_i) = \mathrm{F}_i(h; \alpha_i)$, and we get the following decision-utility

$$u_i^{SA}(h, a_i; \alpha_i) = \mathbb{E}[\pi_i | (h, a_i); \alpha_i] - \theta_i \sum_{j \neq i} \mathrm{F}_i(h; \alpha_i) \mathbb{E}\left[\pi_j | (h, a_i); \alpha_i\right]. \tag{3}$$

We demonstrate SA via the ultimatum minigame in Fig. 1 (Gale et al., 1995), a simplified version of the ultimatum game of Güth et al. (1982). It has a similar structure as the stage game of Selten's (1978) chain-store game. As noted by Gale et al. (1995) (p. 76), models such as ours which imply that players will reject unfair offers "provide a possible resolution of the chain-store paradox that applies even in the case when there is just one potential entrant." Selten also considers that frustration and aggression may be relevant. Vis-á-vis Case 1 in Section 1, the setup can also be interpreted as representing a monopoly seller (Ann) who can offer Bob either a high or a low split of the gains from trade.

Example 1. Ann and Bob (a & b in Fig. 1) negotiate: Ann can make a fair offer f, which is automatically accepted, or a greedy offer g in which case Bob's frustration is

$$\mathrm{F}_b(g; \alpha_b) = [(1 - \alpha_b(g)) \cdot 2 + \alpha_b(g)\alpha_b(y|g) \cdot 1 - 1]^+.$$

Therefore

$$u_b^{SA}(g, n; \alpha_b) - u_b^{SA}(g, y; \alpha_b) = 3\theta_b [2(1 - \alpha_b(g)) + \alpha_b(g)\alpha_b(y|g) - 1]^+ - 1.$$

For Bob to be frustrated he must not expect g with certainty. The less he expects g, and—interestingly—the less he plans to reject, the more prone he is to reject. The more resigned Bob is to getting a low payoff, the less frustrated and prone to aggression he is. Furthermore, it is readily seen how our model can generate non-consequentialist behavior. Holding beliefs and other payoffs constant, increasing Bob's payoff from f will lead to greater frustration after g, and so increases the disutility Bob receives from Ann's material payoff. This makes rejection (punishment of Ann) more attractive. □

In the example, Bob rejects because he is truly angry and prefers n to y. He is not signaling his type to deter Ann from choosing g in the future. This marks a difference with "reputational models." In the example, there is no future behavior for Bob to influence.

Under SA a frustrated player goes after others rather indiscriminately. The word "rather" is justified because, as regards targets of aggression, our modeling of SA restricts attention to co-players, implicitly saying that persons who are not represented in a game are not targets. So the modeler has a responsibility to represent the appropriate environment. If another individual is a relevant target (*e.g.*, Bob's wife in addition to Ann), that person should be included (*e.g.*, as a dummy player). The exact determination of whom to include in the games is an empirical question. We have the qualitative idea that SA allows for innocent targets, and this is what we model. Future research may generate more nuanced insights.

We next consider models where targets of aggression must be less innocent.

Anger from blaming behavior (ABB) Action tendencies may depend on a player's cognitive appraisal of how to blame. When a frustrated player i blames co-players for their *behavior*, he examines the actions chosen in stage 1, without considering others' intentions: How much i blames j is defined by a function that specifies blame $\mathrm{B}_{ij}(a^1; \alpha_i)$ that depends on α_i (but

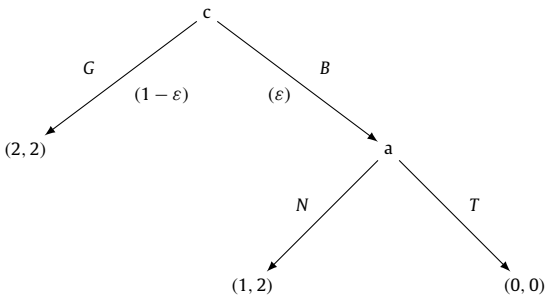

Fig. 2. Hammering one's thumb.

not β_i). Different functional forms are conceivable, but we will here focus on the following one: When frustrated i considers, for each j, what he would have obtained at most, in expectation, had j chosen differently:

$$\max_{a'_j \in A_j(\varnothing)} \mathbb{E}\left[\pi_i | (a^1_{-j}, a'_j); \alpha_i\right].$$

If this could-have-been payoff is more than what i currently expects (that is, $\mathbb{E}[\pi_i | a^1; \alpha_i]$), then i blames j, up to i's frustration (so (2) holds):

$$\mathrm{B}_{ij}(a^1; \alpha_i) = \min\left\{\left[\max_{a'_j \in A_j(\varnothing)} \mathbb{E}\left[\pi_i | (a^1_{-j}, a'_j); \alpha_i\right] - \mathbb{E}[\pi_i | a^1; \alpha_i]\right]^+, \mathrm{F}_i(a^1; \alpha_i)\right\}. \tag{4}$$

Note that j cannot be blamed if he is not active in the first stage. With this, i's decision utility with **anger from blaming behavior (ABB)** is

$$u_i^{ABB}(h, a_i; \alpha_i) = \mathbb{E}\left[\pi_i | (h, a_i); \alpha_i\right] - \theta_i \sum_{j \neq i} \mathrm{B}_{ij}(h; \alpha_i) \mathbb{E}\left[\pi_j | (h, a_i); \alpha_i\right].$$

The essence of Example 1 can also be demonstrated with ABB instead of SA, but the next example exhibits a difference between the two approaches:

Example 2. (Inspired by Frijda, 1993) Consider Fig. 2. Andy the handyman (a) uses a hammer. His apprentice, Bob (b), is inactive. On a bad day (determined by chance, c) Andy hammers his thumb and can then take (T) it out on Bob, or not (N). Assuming $\alpha_a(B) = \varepsilon < 1/2$, we have

$$\mathrm{F}_a(B; \alpha_a) = (1 - \varepsilon) \cdot 2 + \varepsilon \alpha_a(N|B) \cdot 1 - 1 > 0.$$

With SA and with θ_a sufficiently high, on a bad day Andy chooses T in a fit of displaced aggression. But, since Bob is passive, with ABB Andy chooses N regardless of θ_a. □

Next, we consider a more nuanced notion of blame, where players are concerned with their co-players' intentions, and preferences therefore depend upon second-order beliefs.

Anger from blaming intentions (ABI) A player i prone to **anger from blaming intentions (ABI)** asks himself, for each $j \neq i$, whether j intended to give i a low expected payoff. Since such intention depends on α_j (which include j's plan, $\alpha_{j,j}$), how much i blames j depends on i's second-order beliefs β_i, and the decision utility function has the form (1).

The maximum payoff that j, initially, can expect to give to i is

$$\max_{a^1_j \in A_j(\varnothing)} \sum_{a^1_{-j} \in A_{-j}(\varnothing)} \alpha_{j,-j}(a^1_{-j}) \mathbb{E}\left[\pi_i | \left(a^1_j, a^1_{-j}\right); \alpha_j\right]$$
$$\geq \sum_{a^1 \in A(\varnothing)} \alpha_j(a^1) \mathbb{E}\left[\pi_i | a^1; \alpha_j\right] = \mathbb{E}\left[\pi_i | \alpha_j\right],$$

where the inequality holds by definition and the equality is implied by the chain rule. Note also that $\alpha_j(\cdot|a^1)$ is kept fixed under the maximization; we focus on what j initially believes he could achieve, taking the view that at the root he cannot control a^2_j but predicts his choice in stage 2. We assume that i's blame of j at a^1 equals i's expectation, given β_i and conditional on a^1, of the difference between the maximum payoff that j can expect to give to i and what j actually plans/expects to give to i, capped by i's frustration:

$$\mathrm{B}_{ij}(a^1;\beta_i)=\min\left\{\mathbb{E}\left[\max_{a^1_j}\sum_{a^1_{-j}}\alpha_{j,-j}(a^1_{-j})\mathbb{E}\left[\pi_i|\left(a^1_j,a^1_{-j}\right);\alpha_j\right]-\mathbb{E}[\pi_i;\alpha_j]\,\middle|\,a^1;\beta_i\right],\mathrm{F}_i(a^1;\alpha_i)\right\},\tag{5}$$

where α_i is derived from β_i. The expression is nonnegative as per the previously highlighted inequality. Now, using (5), i's decision utility after $h=a^1$ is

$$u_i^{ABI}(h,a_i;\beta_i)=\mathbb{E}\left[\pi_i|(h,a_i);\alpha_i\right]-\theta_i\sum_{j\neq i}\mathrm{B}_{ij}(h;\beta_i)\mathbb{E}\left[\pi_j|(h,a_i);\alpha_i\right].$$

We are ready to illustrate a difference between ABI and each of SA and ABB:

Example 3. Return to Fig. 1. The maximum Ann can expect to give Bob is 2, independently of α_a. Suppose Bob, upon observing g, is certain Ann planned to offer g with probability $p<1$: $\beta_b(\alpha_a(g)=p|g)=1$, and that Ann expected him to accept with probability q: $\beta_b(\alpha_a(y|g)=q|g)=1$. Finally, suppose Bob initially expected to get the *fair* offer ($\alpha_b(f)=1$), so his frustration after g is $\mathrm{F}_b(a^1;\alpha_b)=2-1=1$. We get

$$\mathrm{B}_{ba}(g;\beta_b)=\min\{2-[2(1-p)+qp],1\}=\min\{p(2-q),1\}.$$

If p is low enough, or q high enough, unlike the case with SA or ABB, Bob does not blame all frustration on Ann. She gets some credit for her initial intention to choose f with probability $1-p>0$, and the degree of that credit depends on q. □

3. Equilibrium analysis of leader-follower (LF-)games

We now develop and systematically explore predictions for the three models introduced in Section 2 (SA, ABB, ABI). Our main focus will be on leader-follower games with perfect information, a class rich enough to highlight striking behavioral patterns yet narrow enough to permit sharp results that games outside that class defy. There are two players, and only one is active in each stage. In the first stage the leader (denoted by ℓ) is active. In the second stage, the follower (denoted by f) is active. Thus, ℓ does not move in stage 2, f does not move in stage 1, and there is no third party. Formally:

Definition 1. A game form is called a **leader-follower** (**LF-**)game if $H=\{\varnothing\}\cup A(\varnothing)$, $I=\{\ell,f\}$, $I(\varnothing)=\{\ell\}$, and $I(a^1)=\{f\}$ or $I\left(a^1\right)=\emptyset$ for every $a^1\in A(\varnothing)$.

Condition $H=\{\varnothing\}\cup A(\varnothing)$ here says that no action of the leader terminates the game. This simplifies the exposition and is without loss of generality, because the follower's set of feasible actions may be a singleton after some a^1, in this case $I\left(a^1\right)=\emptyset$. For example, we interpret the Ultimatum Minigame of Fig. 1 as a game form where the responder is forced to accept the fair offer. The (regular) ultimatum game, the chain store game (Selten, 1978), and the "pure threats game" of Klein and O'Flaherty (1993, Fig. 2) are other examples of LF-games. Hammering-one's-thumb (Fig. 2) is not.

Sequential equilibrium (SE) While we depart from traditional game-theoretic analysis in using belief-dependent decision utility, our analysis is otherwise traditional. We adapt B&D's sequential equilibrium (SE) concept.[14] We consider a complete information framework where the rules of the game and players' (psychological) preferences are common knowledge.[15] We interpret an SE as a profile of strategies and beliefs representing a "commonly understood" way to play the game by rational (utility maximizing) agents.[16] The SE concept gives equilibrium conditions for infinite hierarchies of conditional probability systems. In our particular application, utility functions only depend on first- or second-order beliefs, so we define SE for assessments comprising beliefs up to only the second order. Since, technically, first-order beliefs are features of second-order beliefs (see Section 2), we provide definitions that depend only on second-order beliefs, which give SEs for games where psychological utility functions depend only on first-order beliefs as a special case.

Fix a game form (as defined in Section 2, so LF-games are a special case) and decision utility functions $u_i(h,\cdot;\cdot):A_i(h)\times\Delta_i^2\to\mathbb{R}$ ($i\in I$, $h\in H$). This gives a **psychological game** in the sense of B&D (Section 6). An **assessment** is a profile

[14] B&D extend Kreps and Wilson's (1982) classic notion of SE to psychological games; we consider the version (B&D, Section 6) for preferences with own-plan dependence and "local" psychological utility functions. A difference with B&D is that they assume it common knowledge that players behave as planned, whereas we separate plans from behavior and let the consistency of behavior with plan be a rationality condition.

[15] This is partly a modeling choice made for reasons of analytical tractability, but we note that the approach is supported by recent intriguing experimental evidence reported by van Leeuwen et al. (2018). They find that (pre-play) "facial cues provide a credible signal of destructive behavior," and that subjects are to a degree capable of recognizing such "angry buttons." For an analysis of incomplete-information psychological games see Battigalli et al. (2019).

[16] This is a choice of focus more than an endorsement of SE as a solution concept. SE requires that each player i is certain and never changes his mind about the true beliefs and plans, hence intentions, of his co-players. We find this feature questionable. B&D (Sections 2, 5) and Battigalli et al. (2019) argue that, with belief-dependent preferences, alternatives to SE like rationalizability, forward induction, and self-confirming equilibrium are even more plausible than with standard preferences.

of behavior strategies and beliefs $(\sigma_i, \beta_i)_{i\in I} \in \times_{i\in I}(\Sigma_i \times \Delta_i^2)$ such that $\Sigma_i = \times_{h\in H}\Delta(A_i(h))$ and σ_i is the plan $\alpha_{i,i}$ entailed by second-order belief β_i:

$$\sigma_i(a_i|h) = \alpha_{i,i}(a_i|h) = \beta_i\left(Z(h,a_i)\times\Delta^1_{-i}|h\right) \tag{6}$$

for all $i \in I$, $h \in H$, $a_i \in A_i(h)$. Eq. (6) implies that the behavior strategies contained in an assessment are implicitly determined by players' beliefs about paths; therefore, they could be dispensed with. Yet, we follow B&D and make behavior strategies explicit in assessments to facilitate comparisons with the refinements literature.

Definition 2. Assessment $(\sigma_i, \beta_i)_{i\in I}$ is **consistent** if for all $i \in I$, $h \in H$, $a = \left(a_j\right)_{j\in I} \in A(h)$

(a) $\alpha_i(a|h) = \prod_{j\in I}\sigma_j(a_j|h)$,
(b) $\text{marg}_{\Delta^1_{-i}}\beta_i(\cdot|h) = \delta_{\alpha_{-i}}$;

α_i is derived from β_i and $\delta_{\alpha_{-i}}$ is the Dirac measure assigning probability 1 to $\{\alpha_{-i}\} \subseteq \Delta^1_{-i}$.

Condition (a) requires players' beliefs about actions to satisfy independence across co-players (on top of own-action independence), and—conditional on each h—each i expects each j to behave in the continuation as specified by j's plan $\sigma_j = \alpha_{j,j}$, even though j previously deviated from $\alpha_{j,j}$. All players thus have the same first-order beliefs. Condition (b) requires that players' beliefs about co-players' first-order beliefs (hence their plans) are correct and never change, on or off the path. Thus all players, essentially, have the same second-order beliefs (considering that they are introspective and so know their own first-order beliefs). These conditions faithfully capture the "trembling-hand" interpretation of deviations implicit in Kreps and Wilson's (1982) original definition of SE: if i observed deviations from the plans α_{-i} he ascribes to his co-players (according to β_i), instead of changing his mind about the co-players' plans, he would conclude they made mistakes carrying out their plans. Such mistakes are independent across nodes and players, hence the probability of further deviations is negligible.

Definition 3. Assessment $(\sigma_i, \beta_i)_{i\in I}$ is a **sequential equilibrium (SE)** if it is consistent and satisfies the following sequential rationality condition: for all $h \in H$ and $i \in I(h)$, $Supp\sigma_i(\cdot|h) \subseteq \arg\max_{a_i\in A_i(h)} u_i(h, a_i; \beta_i)$.

It can be checked that this definition is equivalent to the traditional one if players have standard preferences, *i.e.*, with a profile of utility functions $\left(v_i : Z \to \mathbb{R}\right)_{i\in I}$ such that $u_i(h, a_i; \beta_i) = \mathbb{E}[v_i|(h, a_i); \alpha_i]$.[17] A special case is the **material-payoff game**, where $v_i = \pi_i$ for each $i \in I$.

Remark 3. Every material-payoff game with *perfect information* and *no relevant ties* has a unique SE, which is in pure strategies and can be computed by backward induction.

As is known from previous work, with psychological utilities uniqueness of equilibrium with deterministic plans may fail even in game forms with perfect information and with no relevant ties.[18] The examples in Section 3 illustrate this for the case of anger-prone players. On the other hand, in our framework existence of equilibrium with (possibly) non-deterministic plans is guaranteed under mild conditions.

Theorem 1. *If $u_i(h, a_i; \cdot)$ is continuous for all $i \in I$, $h \in H$ and $a_i \in A_i(h)$, then there is at least one SE.*

B&D prove a version of this result where first-order beliefs are modeled as belief systems over pure strategy profiles. Our setting adds personal histories, and here first-order beliefs are modeled as beliefs about paths. Given those modifications, the "trembling-hand" technique used in B&D's Theorem 9 can be applied to establish the result. We omit the details.[19]

What we said so far about SE does not assume specific functional forms. From now on, we focus on u_i^{SA}, u_i^{ABB}, and u_i^{ABI}. Since frustration and blame are continuous in beliefs, decision utility is also continuous, and we obtain existence in all cases of interest:

Theorem 2. *Every game with SA, ABB, or ABI has at least one SE.*

[17] According to the standard definition of SE, sequential rationality is given by global maximization over (continuation) strategies at each $h \in H$. By the One-Shot-Deviation principle, in the standard case this is equivalent to "local" maximization over actions at each $h \in H$.

[18] See Geanakoplos et al. and B&D.

[19] A similar technique is used in the proof of Proposition 2 (first part) in the appendix.

Properties of SE Next, we state two general results that we will later apply to LF-games. First, in pure-strategy SE, players are never frustrated on the equilibrium path:

Proposition 1. *Let $(\sigma_i, \beta_i)_{i\in I}$ be an SE assessment of a game with SA, ABB, or ABI; if a history $h \in H$ has probability 1 under profile $(\sigma_i)_{i\in I}$, then*

$$\mathrm{F}_i(h';\alpha_i) = 0 \text{ and } \mathrm{Supp}\sigma_i(\cdot|h') \subseteq \arg\max_{a_i'\in A_i(h')} \mathbb{E}[\pi_i|\left(h', a_i'\right);\alpha_i]$$

for all $h' \preceq h$ and $i \in I$, where α_i is derived from β_i. Therefore, an SE strategy profile of a game with SA, ABB, or ABI with randomization (if any) only in the last stage is also a Nash equilibrium of the agent form of the corresponding material-payoff game.

To illustrate in an LF-game, in Fig. 1, (f, n) can be an SE strategy pair under ABB, and is a Nash equilibrium of the agent form with material-payoff utilities.[20] With (counterfactual) anger, n becomes a credible threat. Note also that Proposition 1 implies that anger is behaviorally irrelevant in one-stage, simultaneous-move games: in our approach, frustration must be triggered by another player's previous action if it is to influence behavior.

Second, recall that two assessments are **realization-equivalent** if the corresponding strategy profiles yield the same probability distribution over terminal histories. The next result demonstrates that the material-payoff SE, with concordant beliefs, is also an equilibrium with belief-dependent anger. We have:

Proposition 2. *In every perfect-information (two-stage) game form with no chance moves and no relevant ties, the unique material-payoff equilibrium is realization-equivalent to an SE of the psychological game with ABI, ABB, or—with only two players—SA.*

Material-payoff SEs of perfect-information games (including LF-games) with no relevant ties are unique and must be in pure strategies (see Remark 3). By Proposition 1, players must maximize their material payoff on the path even if they are prone to anger. As for off-equilibrium path decision nodes, deviations from the material-payoff SE strategies can only be due to the desire to hurt the first-mover, which can only increase his incentive to stick to the material-payoff SE action.

Let us next illustrate how the SE concept works under SA, ABB, and ABI in our favorite LF-example, the ultimatum minigame:

Example 4. Consider Fig. 1. By Proposition 2, every utility function discussed admits the material-payoff SE (g, y) as an SE, regardless of anger sensitivity. To check this, just note that, if Bob expects g, he cannot be frustrated, so—when asked to play—he maximizes his material payoff. Under SA and ABB, (f, n) qualifies as another SE if $\theta_b \geq 1/3$; following g, Bob would be frustrated and choose n, so Ann chooses f. Under ABI (f, n) cannot be an SE. To verify, assume it were, so $\alpha_a(f) = 1$. Since the SE concept does not allow for players revising beliefs about beliefs, we get $\beta_b(\alpha_a(f) = 1|g) = 1$ and $\mathrm{B}_{ba}(g;\beta_b) = 0$; Bob maintains his belief that Ann planned to choose f, hence she intended to maximize Bob's payoff. Hence, Bob would choose y, contradicting that (f, n) is an SE. Next, note that (g, n) is not an SE under any concept: Given SE beliefs Bob would not be frustrated and hence would choose y. The only way to observe rejected offers with positive probability in an SE is with non-deterministic plans. To find such an SE, note that we need $\alpha_a(g) \in (0, 1)$; if $\alpha_a(g) = 0$ Bob would not be reached and if $\alpha_a(g) = 1$ he would not be frustrated, and hence, he would choose y. Since Ann uses a non-degenerate plan she must be indifferent, so $\alpha_b(y) = 2/3$, implying that Bob is indifferent too. In SE, Bob's frustration at g is $\left[2\left(1-\alpha_a(g)\right) + \frac{2}{3}\alpha_a(g) - 1\right]^+ = \left[1 - \frac{4}{3}\alpha_a(g)\right]^+$, which equals his blame of Ann under SA and ABB. Hence we get the indifference condition

$$1 - \theta_b\left[1 - \frac{4}{3}\alpha_a(g)\right]^+ \cdot 3 = 0 - \theta_b\left[1 - \frac{4}{3}\alpha_a(g)\right]^+ \cdot 0$$

$$\Longleftrightarrow$$

$$\alpha_a(g) = \frac{3}{4} - \frac{1}{4\theta_b},$$

where $\theta_b \geq 1/3$. The higher is θ_b the more likely Bob is to get the low offer, so Bob's initial expectations, and hence his frustration and blame, is kept low. Under ABI we get another indifference condition:

[20] In the agent form of a game, each h where player i is active corresponds to a copy (i, h) of i with strategy set $A_i(h)$ and the same utility function as i.

$$1-\theta_b \mathrm{B}_{ba}(g;\beta_b)\cdot 3 = 0-\theta_b \mathrm{B}_{ba}(g;\beta_b)\cdot 0$$

$$\Longleftrightarrow$$

$$1-\theta_b \min\left\{\frac{4}{3}\alpha_a(g), \left[1-\frac{4}{3}\alpha_a(g)\right]^+\right\}\cdot 3 = 0.$$

The right term in braces is Bob's frustration, while

$$\frac{4}{3}\alpha_a(g) = 2-\left[2(1-\alpha_a(g))+\frac{2}{3}\alpha_a(g)\right]$$

is the difference between the maximum payoff Ann could plan for Bob and the actual one. The first term is lower if $\alpha_a(g)\geq 3/8$; if, with that, we can solve the equation, we duplicate the SA/ABB-solution; this is doable if $\theta_b > 1/3$. If $\theta_b \geq 2/3$, with ABI, there is a second non-degenerate equilibrium plan with $\alpha_a(g)\in(0,\frac{3}{8})$ where $\alpha_a(g) = 1/(4\theta_b)$; to see this, solve the ABI indifference condition assuming $\frac{4}{3}\alpha_a(g)\leq 1-\frac{4}{3}\alpha_a(g)$. This SE exhibits starkly different comparative statics: The higher is θ_b, the less likely Bob is to get a low offer and the less he blames Ann following g in light of her intention to choose f with higher probability. □

In the example, the set of SE under ABI may differ from the set of SE under SA and ABB which, however, agree. This is no fluke. In LF-games, under ABB the follower's frustration can only be blamed on the leader, so in these games SA and ABB are equivalent. Let us write u_{i,θ_i} to make the dependence of u_i on θ_i explicit. We can thus note:

Remark 4. In LF-games, SA and ABB coincide, i.e., $u^{SA}_{i,\theta_i} = u^{ABB}_{i,\theta_i}$ for all θ_i.

This property is not guaranteed for non-LF-games, as the following two examples show:

Example 5. Consider the extension of the ultimatum minigame in Fig. 3. It adds a third player, Darryl (d, because we leave c for chance), whose payoffs are represented by the third element in the payoff vectors. It also adds an alternative way for Bob to accept by choosing y' which punishes Darryl by $x>0$. If Ann chooses g, assuming $\alpha_b(g)=\varepsilon<\frac{1}{2}$, we have

$$\mathrm{F}_b(g;\alpha_b) = (1-\varepsilon)\cdot 2+\varepsilon\Big(\alpha_b(y|g)+\alpha_b(y'|g)\Big)\cdot 1-1>0.$$

With SA and $\theta_b>0$, if $x>3$, after g Bob chooses y' in another example of displaced aggression. But, since Darryl is passive and does not move in the first stage, with ABB Bob does not blame Darryl, and Bob is indifferent between y and y', regardless of θ_b. Note that this implies that with SA Bob does not choose n in any SE, while with ABB the pure-strategy SE (f,n) remains. □

LF-games (as in Definition 1) have a single follower, and Example 5 shows that the statement in Remark 4 does not extend to games with more followers. LF-games also do not allow for chance moves, and our next example shows that their presence may impede other attempts at extending Remark 4:

Example 6. Consider again the hammering-ones-thumb game (Fig. 2). With u^{ABB}_a (either version), or u^{ABI}_a, Andy will not blame Bob so his SE-choice is N. But with u^{SA}_a Andy may choose T. Recall that $\mathrm{F}_a(B;\alpha_a)=2(1-\varepsilon)+\varepsilon\alpha_a(N|B)-1$, so the more likely Andy believes it to be that he will take it out on Bob, the less he expects initially and the less frustrated he is after B. Yet, in SE, the higher is θ_a the more likely Andy is to take it out on Bob: Andy's utility from N and T is

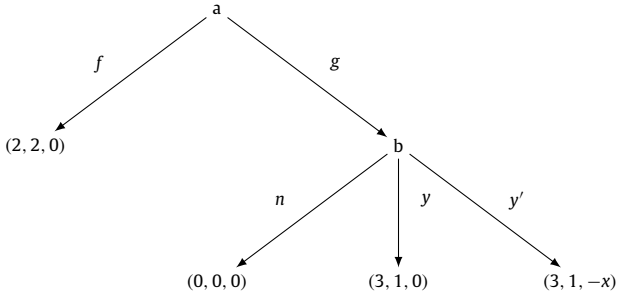

Fig. 3. Ultimatum minigame with a bystander.

$$u_a^{SA}(B,N;\alpha_a) = 1-\theta_a[2(1-\varepsilon)+\varepsilon\alpha_a(N|B)-1]\cdot 2,$$
$$u_a^{SA}(B,T;\alpha_a) = 0-\theta_a[2(1-\varepsilon)+\varepsilon\alpha_a(N|B)-1]\cdot 0 = 0.$$

Sequential rationality of SE implies that one possibility is $\alpha_a(N|B)=1$ and $u_a^{SA}(B,N;\alpha_a)\geq u_a^{SA}(B,T;\alpha_a)$, implying $\theta_a \leq \frac{1}{2(1-\varepsilon)}$. Another possibility is $\alpha_a(N|B)=0$ and $u_a^{SA}(B,N;\alpha_a)\leq u_a^{SA}(B,T;\alpha_a)$, implying $\theta_a \geq \frac{1}{2(1-2\varepsilon)}$. I.e., if Andy is sufficiently susceptible to SA, on bad days he takes his frustration out on Bob. If $\theta_a \in (\frac{1}{2(1-\varepsilon)}, \frac{1}{2(1-2\varepsilon)})$, we can solve for the unique SE in which $u_a^{SA}(B,N;\alpha_a)=u_a^{SA}(B,T;\alpha_a)$ and $\alpha_a(N|B)=\frac{1}{2\varepsilon\theta_a}-\frac{1-2\varepsilon}{\varepsilon}\in(0,1)$. □

If $\theta_a \in (\frac{1}{2(1-\varepsilon)}, \frac{1}{2(1-2\varepsilon)})$, Example 6 additionally shows how we cannot, in general, take for granted the existence of an SE with a deterministic path (a point relevant also for u_i^{ABB}or u_i^{ABI} in other games).[21] In LF-games with no relevant ties that point is moot, however. To see this, note that such games are covered by Remark 3 and Proposition 2. These results imply that LF-games with no relevant ties have a unique material-payoff SE in pure strategies that is realization-equivalent to an SE with SA, ABB, or ABI.

Example 6 also illustrates how we cannot, in general, take for granted the existence of an SE in which no player is frustrated along the path of play. In LF-games with no relevant ties, also that point is moot, because in the SE with SA, ABB, or ABI realization-equivalent to the material-payoff SE the path is deterministic and the follower cannot be frustrated on path.

In LF-games, our model captures an important aspect of the psychology of anger. Sell et al. (2009) argue that anger "is produced by a neurocognitive program engineered by natural selection to use bargaining tactics to resolve conflicts of interest in favor of the angry individual." Our approach is consistent with this view. First note that our complete-information analysis assumes that the leader knows how anger-prone the follower is. This is a good approximation for interactions between agents who know each other well, and also a good approximation of face-to-face interactions, given that humans are good at reading facial cues to infer personality traits such as trait-anger (van Leeuwen et al., 2018). With this, our models predict that in LF-game forms anger-prone followers will obtain at least as large a material payoff as self-interested ones. In our evolutionary past, higher material payoffs meant longer survival and better access (for males) to sexual partners, both of which yield higher reproduction rates (*e.g.*, Buss, 2016). Hence, we argue that our approach supports the claim of Sell et al.

Proposition 3. *In every LF-game with no relevant ties, the expected material payoff of the follower in any SE with SA, or ABB, or ABI is at least as large as the material payoff of the follower in the unique material-payoff SE.*

To see why this is the case, consider an SE of the psychological game that yields a different expected material payoff to the follower than the material-payoff equilibrium. Since a first mover cannot be frustrated, all the actions chosen by the leader with positive probability in this SE must maximize his expected material payoff. Thus, if in this SE the leader deviates from the material-payoff equilibrium action, it must be the case that he correctly expects this action to be "punished" by the follower with a deviation from the follower's material-payoff maximizing reply. This in turn requires that the follower is frustrated by the material-payoff equilibrium action of the leader, hence, that his expected payoff in this SE is higher than in the material-payoff equilibrium.

In game forms with more than two followers, Proposition 3 need not hold. For example, there may exist equilibria where one follower behaves as-if self-interested, while the other is frustrated by the material-payoff equilibrium outcome. This can result in the first follower getting less than in the material-payoff equilibrium as shown by the following example.

Example 7. Consider the game in Fig. 4. The leader, Ann, chooses between Left (l) and Right (r). In the left subgame Bob chooses between the payoff profile $(7,1,3)$ and 0 for all players. In the right subgame Darryl chooses between the payoff profile $(6,2,2)$ and 0 for all players. The material-payoff SE is (l,y,y') and in this equilibrium Darryl gets 3. However, with SA, ABB, or ABI, for sufficiently large values of θ_b and for all $\theta_d \geq 0$, the strategy profile (r,n,y') can also be an SE. To see this note that if Ann deviates to l, Bob will be frustrated, and if θ_b is large enough Bob will choose n after l. In this equilibrium Darryl gets 2, a payoff smaller than in the material payoff equilibrium. □

Next, we demonstrate that our models of anger are behaviorally relevant only in the subclass of LF-games involving threats. As above, consider a LF-game with no relevant ties. In these, the leader has a unique best response to each pure strategy $s_f \in \times_{a_\ell \in A_\ell(\varnothing)} A_f(a_\ell)$ of the follower. We let

$$r_\ell(s_f) = \arg\max_{a_\ell \in A_\ell(\varnothing)} \pi_\ell(a_\ell, s_f(a_\ell))$$

[21] In the example it even happens with a single active player, highlighting that we deal with a psychological game, as this could never be the case in a standard game.

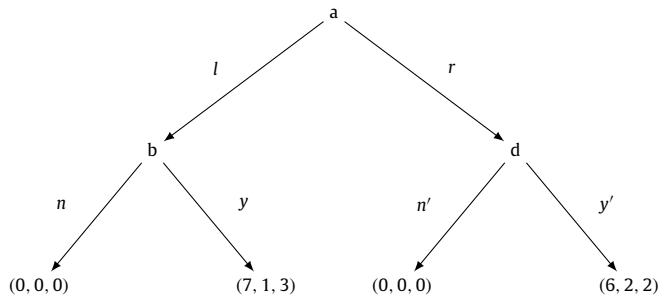

Fig. 4. A game with two followers.

denote this best response. Let $(\bar{a}_\ell, \bar{s}_f) = (r_\ell(\bar{s}_f), \bar{s}_f)$ denote the unique material-payoff SE, where $\bar{s}_f$ is the material equilibrium pure strategy of the follower. With this, we define a threat as a strategy that penalizes the leader for playing the material-payoff SE:

Definition 4. In any LF-game with *no relevant ties*, a **threat** of player f is a pure strategy $\hat{s}_f$ such that, relative to the material-payoff SE:

1. Implementing the threat harms the leader: $\pi_\ell(\bar{a}_\ell, \hat{s}_f(\bar{a}_\ell)) < \pi_\ell(\bar{a}_\ell, \bar{s}_f(\bar{a}_\ell))$.
2. The leader's best response to the threat strategy benefits the follower (who subsequently maximizes his material payoff): let $\hat{a}_\ell = r_\ell(\hat{s}_f)$, then $\pi_f(\hat{a}_\ell, \hat{s}_f(\hat{a}_\ell)) = \pi_f(\hat{a}_\ell, \bar{s}_f(\hat{a}_\ell)) > \pi_f(\bar{a}_\ell, \bar{s}_f(\bar{a}_\ell))$.

Our definition of a threat is inspired by, but differs from, that of Klein and O'Flaherty (1993). Note that the threat must be costly to implement along the path of the material payoff SE, because — by condition 1 — it differs from the unique material best response. Thus, the above conditions incorporate the notion of a threat that would not be credible if the follower were known to be a material payoff maximizer. The paradigmatic example of a game form with threats is the Ultimatum Game: indeed, the strategy n of rejecting the greedy offer in the Ultimatum Minigame of Fig. 1 is a threat. If the leader best responds to the threat, the resulting strategy pair is a Nash equilibrium of the material payoff game. This is a general property of threats according to Definition 4: indeed, $\pi_\ell(a_\ell, \bar{s}_f(a_\ell)) = \max_{a_f \in A_f(a_\ell)} \pi_f(a_\ell, a_f)$ because $\bar{s}_f$ is the material-payoff backward induction strategy, and every strategy pair $(\hat{a}_\ell, \hat{s}_f)$ such that $\hat{a}_\ell = r_\ell(\hat{s}_f)$ and $\pi_f(\hat{a}_\ell, \hat{s}_f(\hat{a}_\ell)) = \max_{a_f \in A_f(\hat{a}_\ell)} \pi_f(\hat{a}_\ell, a_f)$ is a material-payoff Nash equilibrium. Furthermore, we note that we made Condition 1 in the definition explicit for conceptual clarity, although it is implied by Condition 2 and the assumption of no relevant ties. To see this, note that the inequality in Condition 2 implies that $\bar{a}_\ell \neq \hat{a}_\ell$. By no relevant ties, $\hat{a}_\ell$ (respectively, $\bar{a}_\ell$) is the *unique* best reply to $\hat{s}_f$ (respectively, $\bar{s}_f$), and equality $\pi_f(\hat{a}_\ell, \hat{s}_f(\hat{a}_\ell)) = \pi_f(\hat{a}_\ell, \bar{s}_f(\hat{a}_\ell))$ in Condition 2 implies $\hat{s}_f(\hat{a}_\ell) = \bar{s}_f(\hat{a}_\ell)$. Therefore,

$$\pi_\ell(\bar{a}_\ell, \hat{s}_f(\bar{a}_\ell)) < \pi_\ell(\hat{a}_\ell, \hat{s}_f(\hat{a}_\ell)) = \pi_\ell(\hat{a}_\ell, \bar{s}_f(\hat{a}_\ell)) < \pi_\ell(\bar{a}_\ell, \bar{s}_f(\hat{a}_\ell)).$$

To summarize:

Remark 5. In every LF-game with no relevant ties, for every pure strategy $\hat{s}_f$ of the follower that satisfies Condition 2 of Definition 4, (i) the strategy pair $(r_\ell(\hat{s}_f), \hat{s}_f)$ is a Nash equilibrium of the material payoff game, and (ii) $\hat{s}_f$ is a threat.

We can use Definition 4 to characterize the SE behavior of anger-prone followers:

Proposition 4. *In every LF-game with no relevant ties where the follower has a* threat, *there exists a pure-strategy SE (the "deterrence SE") with SA/ABB (when the anger sensitivity parameter is sufficiently large) such that*

1. *The follower's strategy is a threat.*
2. *The leader does not play his material-payoff SE action.*
3. *The follower's material payoff is strictly greater than in the material-payoff SE.*

The proof of Proposition 4 involves demonstrating that the consistent assessment where the follower plays a given threat (and the leader best responds to the threat) is sequentially rational when the follower is sufficiently prone to anger, and therefore it constitutes an SE with SA and ABB. This construction does not work with ABI for the reasons explained in Example 4 about the Ultimatum Game: under the "trembling-hand" interpretation of deviations inherent in the SE concept, a deviation from the action $\hat{a}_\ell$ that the follower wants to induce is interpreted by him as unintentional, so that he does not blame the leader and the threat is not credible.

We also provide a comparative statics result that relates anger sensitivity to material payoffs. Proposition 4 demonstrates that the deterrence SE in LF-games with threats exists when the anger sensitivity parameter of the follower θ_f is sufficiently large. The proposition below relates the minimal such sensitivity parameter, $\underline{\theta}_f$, to material payoffs:

Proposition 5. *In every LF-game with no relevant ties where the follower has a* threat*, the minimal value of the anger sensitivity of the follower* $\underline{\theta}_f$ *that supports the deterrence SE is*

1. *Increasing in the follower's material payoff from the material-payoff SE and decreasing in the follower's material payoff from the deterrence SE*
2. *Decreasing in the leader's material payoff from the material-payoff SE*

In the deterrence SE the follower's frustration after the leader deviates to the material payoff strategy is the difference between his material payoffs from the deterrence SE and the material-payoff SE. Greater frustration implies that a lower sensitivity to anger is sufficient to support the deterrence SE. In addition the follower's anger (which determines his decision utility) is increasing in the leader's payoff from deviating to the material payoff SE, so increasing this amount reduces the value of $\underline{\theta}_f$ needed to support the deterrence SE.

The next proposition shows that, if the game form does not include a threat, then the follower behaves as-if self-interested in every pure SE:

Proposition 6. *In every LF-game with* no relevant ties *and* without threats*, every pure SE of the psychological game with SA/ABB or ABI is realization equivalent to the material-payoff SE.*

To illustrate, anger is not relevant in the trust game of Berg et al. (1995) (and the mini-trust game in B&D's Fig. 1). In general, behavioral patterns that require players to place positive weight on a co-player's material payoff cannot be explained via anger.

4. Comparison with other models

We now contrast the behavioral predictions resulting from frustration and anger with other models of strategic and social behavior. We first consider distributional preferences, which transform material payoffs at terminal histories but which otherwise retain the standard assumption that choices depend solely on their consequences in terms of material payoffs. We then discuss models of reciprocity, and some alternative approaches to modeling anger.

Distributional preferences Like anger, models of distributional preferences such as inequality aversion (*e.g.*, Fehr and Schmidt, 1999; Bolton and Ockenfels, 1989) predict costly punishment. However, a number of studies demonstrate that the decision to engage in costly punishment depends upon both payoffs reached at terminal histories as well as payoffs from unreached histories.[22] Our approach can capture this non-consequentialist aspect of behavior, while distributional preferences cannot. For example, in the ultimatum minigame of Fig. 1, Bob's decision to reject the greedy offer may depend upon not only the payoffs from accepting or rejecting the offer, but also upon the payoff that Ann could have given Bob had she chosen the fair offer. Holding payoffs and other beliefs constant, our models predict that Bob will be more likely to reject the greedy offer when either he assigns higher probability to receiving the fair one, or when the fair payoff is increased. In general, distributional preferences cannot capture behavioral patterns which depart from consequentialism, while our models can.

Our models assume that players care only about material payoffs and anger, disregarding distributional considerations. We thereby highlight the effects of frustration, blame, and anger on behavior in strategic interaction. However, real-world decision makers may have a mixture of material, distributional, and psychological motivations (*e.g.*, Falk and Fischbacher, 2006). We leave the exploration of mixed-motive concerns for the future.

Reciprocity Negative reciprocity à la Rabin (1993), Dufwenberg and Kirchsteiger (2004), and Falk and Fischbacher (2006) joins anger as a motivation that triggers hostility. Like anger, negative reciprocity can result in non-consequentialist behavior, but anger and negative reciprocity differ in key ways. We sketch a comparison with Dufwenberg and Kirchsteiger's notion of sequential reciprocity equilibrium (SRE; refer to their article for full definitions).

In the hammering-one's-thumb game (Fig. 2), Andy may take it out on Bob if he is motivated by simple anger. Were he motivated by reciprocity, this could never happen: Bob's kindness, since he is a dummy-player, equals 0, implying that Andy chooses as-if selfish. Reciprocity here reflects intuitions similar to the ABI concept, but that analogy only carries so far, as we show next.

Reciprocity also allows for "miserable equilibria," where a player reciprocates expected unkindness before it occurs. For example, in the Ultimatum Minigame of Fig. 1, (g, n) may be a SRE. Ann makes offer g despite believing that Bob will

[22] Brandts and Sola (2001), Charness and Rabin (2002), Nelson (2002), Falk et al. (2003, 2008), Sutter (2007).

reject; given her beliefs about Bob's beliefs, Ann perceives Bob as seeing this coming, which makes him unkind, so she punishes by choosing g. Such self-fulfilling prophecies of destructive behavior have no counterpart under any of our anger notions. Since Ann moves at the root, Remarks 1 and 2 demonstrate that she cannot be frustrated, and hence chooses as-if selfish. By Proposition 1, sacrificing material payoff to harm a co-player never occurs on the path of a pure-strategy SE with ABB, ABI, or (in two-player games) SA.

Alia Card and Dahl (2011) show that reports of domestic abuse go up when football home teams favored to win lose. They argue that this is in line with the theory of expectations-dependent reference points developed by Kőszegi and Rabin (2006, 2007; henceforth K&R). K&R model the loss felt when a player gets less than he expected, which one may think of as a form of disappointment with negative valence (cf. Bell, 1985, Loomes and Sugden, 1986). However, K&R do not model other-regarding preferences directly: they focus on the consequences of their model for individual decisions. Our models study the social consequences of frustration: frustration results in lower weights on coplayer payoffs, and hence encourages costly punishment. Our simple anger model and the example of hammering-one's-thumb encapsulates Card and Dahl's result. A key difference between this paper and K&R is that in their work anticipation of the negative valence of future frustrations influences decision utility. Our decision makers are influenced by past frustrations, rather than future ones. Important modeling details then distinguish how we define frustration and how K&R define loss (*e.g.*, how we cap frustration using the highest attainable payoff).

In a series of intriguing papers, Rotemberg explores how consumer anger shapes firms' pricing (2005, 2011), as well as interaction in ultimatum games (2008). He proposes (versions of) a theory in which players are slightly altruistic, and consumers/responders also care about their co-players' degrees of altruism. Namely, they abruptly become angry and punish a co-player whom they come to believe has an altruism parameter lower than some threshold. "One can thus think of individual i as acting as a classical statistician who has a null hypothesis that people's altruism parameter is at least as large as some cutoff value. If a person acts so that i is able to reject this hypothesis, individual i gains ill-will towards this person" (Rotemberg, 2008, p. 464). Rotemberg shows how his model impressively captures the action in his data sets. It is natural to wonder whether our approach, which is structured very differently from his (*e.g.*, we make no reference to altruism), could achieve that too. As regards behavior in ultimatum (and some other) games, there is already some existing evidence that is consistent with our approach; see the discussion regarding experiments below. Regarding pricing, we leave for empirical economists the task of exploring the topic.

Winter (2014) and Winter et al. (2016) model anger and other emotions in games with a version of the indirect evolutionary approach[23]: Like us, Winter *et al.* assume that preferences over outcomes are "emotional" and endogenous, but we differ in the way we model emotions and make them endogenous. We assume that emotions depend on endogenous beliefs, while Winter *et al.* model the rest points of an adaptation process of belief-independent preferences.

Brams (2011) studies anger in sequential interactions by modeling players who take turns changing the state of a 2×2 payoff matrix and receive payoffs at the end of the game. However, like Winter's, his model of anger is independent of beliefs, while we argue that beliefs are central to emotions.

Akerlof (2016) models anger in a two-person Bayesian game where the first mover decides whether or not to follow a rule, and the second mover decides whether or not to punish the first mover. Player 1's compliance with the rule is motivated by a sense of duty. Player 2 may be sensitive to anger from noncompliance if she thinks that a "reasonable person," modeled as a person with similar preferences to player 2, would comply. Similarly to Rotemberg, Akerlof motivates costly punishment via preferences over others' types. In contrast, we assume that anger and aggression arise from frustration and payoff expectations. Rotemberg's and Akerlof's approaches begin with norms about behavior, and condition anger on violation of those. In contrast, we develop models that reflect the psychology of frustration and anger.

5. Discussion

Incorporating the effects of emotions in economic analysis is a balancing act. One wants to focus on sentiments that make empirical sense, but human psychology is multi-faceted and there is no unambiguous yardstick. Our formulation provides a starting point for exploring how anger shapes interaction, and experimental or other evidence will help to assess empirical relevance and suggest revised formulas. We conclude by discussing sundry topics that may help gain perspective on, build on, or further develop our work.

Frustration Consider substituting $\mathbb{E}[\pi_i; \alpha_i] - \mathbb{E}[\pi_i | a^1; \alpha_i]$ for $\mathrm{F}_i(a^1; \alpha_i)$ of Section 2. This would measure i's *actual* diminished expectations at a^1, unlike $\mathrm{F}_i(a^1; \alpha_i)$ which reflects diminished expectations relative to what i believes is the most he can get (which we think of as the adequate way to represent the unexpected unavailability of something cared about). To appreciate how dramatically this would impact behavior, consider a two-player common-interest game: Ann chooses *Out* or *In*; in the former case the game ends with payoffs $(1, 1)$, in the latter case Bob chooses between $(0, 0)$ and $(2, 2)$. *Mutatis mutandis*, for high enough θ_b, with the alternative, under SA and ABB, there is an SE where Ann chooses *Out* and Bob would

[23] See, for example, Güth and Kliemt (1998).

go for $(0, 0)$. Following *In*, Bob would be frustrated because he (so-to-say) feels locked-in with his stage-2 planned action. Our formulation of $F_i(a^1; \alpha_i)$ rules that out.

Take a binary gamble where with probability $p > 0$ Ann wins $\$x > 0$, and otherwise gets \$0. Her frustration, using our definition, equals her initial expectation: $p \cdot x$. This embodies strong implications for how frustrations compare across contexts, *e.g.*, the frustration of a highly expected failure to win the state lottery versus that of some unlikely small loss. We are agnostic as regards empirical relevance, but alert the reader to the issue.[24]

The evidence says a player becomes frustrated when his goals are unexpectedly thwarted (see Section 1). Here we focus on material rewards, and Cases 1–3 indicate broad applied potential. Yet one may also imagine other sources of frustration:

Case 4: In 2007 Apple launched its iPhone at \$499. Soon after they introduced a new version at \$399, re-priced the old model \$299, causing outrage among early adopters. Apple paid back the difference. Did this help long run profit?
Case 5: The 2008 TARP bank bail-out infuriated some US voters. Did this ignite Tea Party/Occupy-Wall Street movements?

In case 4, an early adopter is frustrated because he regrets he already bought, not because new information implies his expected rewards drop. In case 5, even an activist who is materially unaffected personally may be frustrated because of unexpected perceived unfairness. These examples are not exhaustive; further sources of frustration may, *e.g.*, involve shocks to self-esteem.[25] Techniques analogous to those we have developed may be applicable in these cases, but going in these directions is left for future research.

As regards the effects of frustration, we considered changes to a player's utility but neglected other plausible adjustments. Gneezy and Imas report data from an intriguing experiment involving two-player zero-sum material payoff games. In one game players gain if they are strong, in another if they are smart. Before play starts, one subject may anger his opponent and force him to stay in the lab to do boring tasks. A thus frustrated player's performance is enhanced when strength is beneficial (possibly from increased adrenaline flow), but reduced when cool logic is needed (as if an angered player becomes cognitively impaired). Our approach can capture the first aspect, but not the second: We can let consequences of actions depend also on beliefs (*e.g.*, because emotions affect strength or speed; cf. Rauh and Seccia, 2006); this ultimately translates into belief-dependent utility (or cost) of actions. However, to model the second effect, we would need a theory of endogenous cognitive abilities.

Valence & action-tendency Psychologists classify emotions in multiple ways. Two prominent aspects are **valence**, the intrinsic pleasantness or aversiveness of an emotion, and **action-tendency**, or how behavior is shaped as the emotion occurs. Both notions have bearing on anger. For example, most psychologists believe anger has negative valence (see, *e.g.*, Harmon-Jones and Sigelman, 2001, p. 978). Perhaps such considerations steer people to avoid frustrations, say by not investing in the stocks. That said, the distinguishing feature of anger psychologists stress concerns its action-tendency of aggression, not its valence. In our theory, we exaggerate this, abstracting away from frustration avoidance, while emphasizing frustration-induced aggression. This is reflected in the decision utility functions, which are shaped by current frustration, not by the anticipation of the negative valence of future frustrations.[26]

Blame We explored various ways a player may blame others, but other notions are conceivable. For example, with anger from blaming behavior i's blame of j depends on what i believes he would truly get at counterfactual histories, rather than the most he could get there. We view this modeling choice as reflecting local agency; i's current agent views other agents of i as uncontrollable, and he has no direct care for their frustrations. Another example relates to how we model anger from blaming intentions: i's blame of j depends on β_i, his second-order beliefs. Recall that the interpretation concerns beliefs about beliefs about material payoffs, not beliefs about beliefs about frustration, which would be third- rather than second-order beliefs. Battigalli and Dufwenberg (2007), in a context which concerned guilt rather than anger, worked with such a notion.

Our blame concepts one way or another assess the *marginal* impact of others. For example, consider a game where i exits a building while all $j \in I \backslash \{i\}$, unexpectedly to i, simultaneously hurl buckets of water at i, who gets soaked. According to our approach, i cannot blame any j as long as there are at least two hurlers. One could imagine that i alternatively blames, say, all the hurlers on the grounds that they *collectively* could thwart i's misery, or that i splits the blame among all hurlers. Halpern (2016, Chapter 6) explores such issues and develops a model that assigns positive blame even when outcomes are overdetermined.

People may also blame others in unfair ways, *e.g.*, nominating scapegoats. Our notions of SA and ABB may embody such notions to some degree. However, it has not been our intention to address this issue systematically.

[24] The example involves just one player, to facilitate frustration-calculation; interesting testable implications obviously arise more generally, *e.g.*, in modified versions of the hammering-one's-thumb game.

[25] See Baumeister et al. (1996) for an interesting discussion linking (threatened) self-esteem and violence.

[26] In previous work we modeled another emotion: guilt (*e.g.*, Battigalli and Dufwenberg, 2007; Chang et al., 2011). To gain perspective note how our approach to valence and action-tendency was then reversed. Guilt has valence (negative!) as well as action-tendency (*e.g.* "repair behavior"; see Silfver, 2007). In modeling guilt we highlighted anticipation of its negative valence while neglecting action-tendency.

Several recent experiments explore interesting aspects of blame (Bartling and Fischbacher, 2012; Gurdal et al., 2014; Celen et al., 2017). Our focus on blame is restricted to its relation to frustration, not on other reasons that may lead people to blame each other.[27]

Anger management People aware of their inclination to be angry may attempt to manage or contain their anger. Our players anticipate how frustrations shape behavior, and they may avoid or seek certain subgames because of that. However, there are interesting related phenomena we do not address: Can i somehow adjust θ_i say by taking an "anger management class?" If so, would rational individuals want to raise, or to lower, their θ_i? How might that depend on the game forms they play and completeness or incompleteness of information? These are potentially relevant questions related to how we have modeled action-tendency. Further issues would arise if we were to consider aspects involving anticipated negative valence of future frustrations, or bursts of anger.

Experiments Our models tell stories of what happens when anger prone players interact. It is natural to wonder about empirical relevance. Experiments may shed light.

A few studies that measure beliefs, emotions, and behavior together provide support for the notion that anger and costly punishment result from outcomes which do not meet expectations. Pillutla and Murningham (1996) find that reported anger predicted rejections better than perceived unfairness in ultimatum games. Fehr and Gächter (2002) elicit self-reports of the level of anger towards free riders in a public goods game, concluding that negative emotions including anger are the proximate cause of costly punishment.

Other studies connect unmet expectations and costly punishment in ultimatum games. Falk et al. (2003) measure beliefs and behavior in ultimatum minigames; higher proportions of rejections of disadvantageous offers when responders' expected payoffs are higher, consistent with our models.[28] Schotter and Sopher (2007) measure second-mover expectations, concluding that unfulfilled expectations drive rejections of low offers. Similarly, Sanfey (2009) finds that psychology students who are told that a typical offer in the ultimatum game is \$4-\$5 reject low offers more frequently than students who are told that a typical offer is \$1–\$2.

A series of papers by Frans van Winden and coauthors records emotions and expectations in power-to-take games (which resemble ultimatum games, but allows for partial rejections).[29] Second-mover expectations about first-mover "take rates" are key in the decision to destroy income, and anger-like emotions are triggered by the difference between expected and actual take rates. The difference between actual and reported "fair" take rate is not significant in determining anger, suggesting that deviations from expectations, rather than from fairness benchmarks, drive anger and the destruction of endowments.

A literature in neuroscience connects expectations with social norms to study the neural underpinnings of emotional behavior. In Xiang et al. (2013), subjects respond to a sequence of ultimatum game offers whilst undergoing brain imaging. Unbeknownst to subjects, the experimenter controls the distribution of offers in order to manipulate beliefs. Rejections occur more often when subjects expect higher offers. The authors connect norm violations (i.e., lower than expected offers) with reward prediction errors from reinforcement learning, which are known to be the computations instantiated by the dopaminergic reward system. Xiang et al. note that "when the expectation (norm) is violated, these error signals serve as control signals to guide choices. They may also serve as the progenitor of subjective feelings."

It would be useful to develop tests specifically designed to target key features of our theory. For example, which version—SA, ABB, ABI—seems more empirically relevant, and how does the answer depend on context (*e.g.*, is SA more relevant for tired subjects)? Some insights may again be gleaned from existing studies. For example, Gurdal et al. (2014) study games where an agent invests on behalf of a principal, choosing between a safe outside option and a risky alternative. If the latter is chosen, then it turns out that many principals punish the agent if and only if by chance a poor outcome is realized. This seems to indicate some relevance of our ABB solution (relative to ABI). That said, Gurdal et al.'s intriguing design is not tailored to specifically test our theory (and beliefs and frustrations are not measured).

A few recent studies are directly motivated by our work. Persson (2018) presents a test of simple anger. He explores the hammering-one's-thumb game, and documents that frustrations occur much as predicted, but no punishments occur. His results thus favor ABB or ABI over SA. More recently, we have ourselves begun to study our models in the laboratory: Aina et al. (2018) devise tests that manipulate the responder's payoff from the proposer's outside option in ultimatum minigames. Dufwenberg et al. (2018a, 2018b) test predictions that link anger to communicated promises and threats.

Extensions & applications We have (mostly) limited attention to psychological LF-games with complete information. An important task for future work is to explore whether and how frustration & anger may matter in other games, *e.g.*, where many players move simultaneously in the first stage, with multiple stages, or with incomplete information. In addition, one may want to explore other solution concepts than SE.

[27] For example, Celen et al. (2017) present a model where i asks how he would have behaved had he been in j's position and had j's beliefs. Then i blames j if j appears to be less generous to i than i would have been, and may blame j even if i is not surprised/frustrated. Or imagine a model where players blame those considered unkind, as defined in reciprocity theory, independently of frustration.

[28] See the data in Fig. 2 and Table 1 of their paper. Brandts and Sola (2001) find similar behavioral results (see Table 1). Compare also with our discussion of Example 3 in Section 2.

[29] Bosman and van Winden (2002), Bosman et al. (2005), Reuben and van Winden (2008).

Formulating, motivating, and elucidating the key definitions of our models has already been more than a mouthful. We have therefore also not taken this paper far in the direction of doing applied economics. Make no mistake about it though, the hope that our models will prove useful for such work has been a primary driving force. Our psychologically grounded models of frustration, anger, and blame may shed light on many of the themes (e.g. pricing, violence, politics, recessions, haggling, terror, and traffic) that we listed at the start of this paper. We hope to do some work in these directions ourselves.

Appendix A

Below we define belief spaces and provide proofs of the propositions.

A.1. Preliminaries

To ease exposition, some key definitions/equations from the main text are repeated below.

For each topological space X, we let $\Delta(X)$ denote the space of Borel probability measures on X endowed with the topology of weak convergence of measures. Every Cartesian product of topological spaces is endowed with the product topology. A topological space X is metrizable if there is a metric that induces its topology. A Cartesian product of a countable (finite, or denumerable) collection of metrizable spaces is metrizable.

The space of first-order beliefs of player i is $\Delta_i^1 \subseteq \times_{h_i \in H_i} \Delta(Z(h_i))$, consisting of all conditional probability systems $\alpha_i = (\alpha_i(\cdot|Z(h_i)))_{h_i \in H_i}$ such that:

- the chain rule holds: for all $h_i, h_i' \in H_i$, if $h_i \prec h_i'$ then for every $Y \subseteq Z(h_i')$

$$\alpha_i(Z(h_i')|Z(h_i)) > 0 \Rightarrow \alpha_i\left(Y|Z(h_i')\right) = \frac{\alpha_i(Y|Z(h_i))}{\alpha_i(Z(h_i')|Z(h_i))}; \tag{7}$$

- i's beliefs satisfy an own-action independence property: for all $h \in H$, $a_i \in A_i(h)$, $a_{-i} \in A_{-i}(h)$ (using obvious abbreviations)

$$\alpha_{i,-i}(a_{-i}|h) = \alpha_{i,-i}(a_{-i}|h, a_i). \tag{8}$$

The space of second-order beliefs of i is $\Delta_i^2 \subseteq \times_{h_i \in H_i} \Delta\left(Z(h_i) \times \Delta_{-i}^1\right)$ – where $\Delta_{-i}^1 = \times_{j \neq i} \Delta_j^1$ – consisting of all conditional probability systems $\beta_i = (\beta_i(\cdot|h_i))_{h_i \in H_i}$ such that:

- the chain rule holds: if $h_i \prec h_i'$ then

$$\beta_i(h_i'|h_i) > 0 \Rightarrow \beta_i\left(E|h_i'\right) = \frac{\beta_i(E|h_i)}{\beta_i(h_i'|h_i)} \tag{9}$$

  for all $h_i, h_i' \in H_i$ and every event $E \subseteq Z(h_i') \times \Delta_{-i}^1$;
- i's second-order beliefs satisfy an own-action independence property:

$$\beta_i\left(Z\left(h, (a_i, a_{-i})\right) \times E_\Delta | (h, a_i)\right) = \beta_i\left(Z\left(h, (a_i', a_{-i})\right) \times E_\Delta | (h, a_i')\right), \tag{10}$$

  for all $h \in H$, $a_i, a_i' \in A_i(h)$, $a_{-i} \in A_{-i}(h)$, and (measurable) $E_\Delta \subseteq \Delta_{-i}^1$.

Note that (8) and (10) are given by equalities between marginal measures (on $A_{-i}(h)$ and $A_{-i}(h) \times \Delta_{-i}^1$ respectively).

Lemma 1. *For each player $i \in I$, Δ_i^2 is a compact metrizable space.*

Proof. Let Θ be a non-empty, compact metrizable space. Lemma 1 in Battigalli and Siniscalchi (1999) (B&S) establishes that the set of arrays of probability measures $(\mu(\cdot|h_i))_{h_i \in H_i} \in \times_{h_i \in H_i} \Delta(Z(h_i) \times \Theta)$ such that

$$h_i \prec h_i' \wedge \mu(h_i'|h_i) > 0 \Rightarrow \mu\left(E|h_i'\right) = \frac{\mu(E|h_i)}{\mu(h_i'|h_i)}$$

is closed. Note that, in the special case where Θ is a singleton, each $\Delta(Z(h_i) \times \Theta)$ is isomorphic to $\Delta(Z(h_i))$; hence, the set of first-order beliefs satisfying (7) is closed. Letting $\Theta = \Delta_{-i}^1$, we obtain that the set of second-order beliefs satisfying the chain rule (9) is closed. Since $\times_{h_i \in H_i} \Delta(Z(h_i))$ is a compact subset of a Euclidean space and eq. (8) is a closed condition (equalities between marginal measures are preserved in the limit), Lemma 1 in B&S implies that Δ_i^1 is a closed subset of a compact metrizable space. Hence, Δ_i^1 is a compact metrizable space.

It is well known that if $X_1, \ldots, X_K$ are compact metrizable, so is $\times_{k=1}^K \Delta(X_k)$ (see Aliprantis and Border, 2006, Theorem 15.11). Hence, by Lemma 1 in B&S, the set of second-order beliefs satisfying (9) is a closed subset of a compact metrizable space. Since eq. (10) is a closed condition, this implies that Δ_i^2 is compact metrizable. □

Lemma 2. *For each profile of behavioral strategies $\sigma = (\sigma_i)_{i\in I}$ there is a unique profile of second-order beliefs $\beta^\sigma = \left(\beta_i^\sigma\right)_{i\in I}$ such that (σ, β^σ) is a consistent assessment. The map $\sigma \mapsto \beta^\sigma$ is continuous.*

Proof. Write $\mathbb{P}^\sigma(h'|h)$ for the probability of reaching h' from h, *e.g.*,

$$\mathbb{P}^\sigma(a^1, a^2|\varnothing) = \left(\prod_{j\in I}\sigma_j(a_j^1|\varnothing)\right)\left(\prod_{j\in I}\sigma_j(a_j^2|a^1)\right).$$

Define $\alpha_i^\sigma(z|h) = \mathbb{P}^\sigma(z|h)$ for all $i \in I$, $h \in H$, $z \in Z$. Define β_i^σ as $\beta_i^\sigma(\cdot|h) = \alpha_i^\sigma(\cdot|h) \times \delta_{\alpha_{-i}^\sigma}$ for all $i \in I$, $h \in H$. It can be checked that (1) $\beta_i^\sigma \in \Delta_i^2$ for each $i \in I$, (2) (σ, β^σ) is a consistent assessment, and (3) if $\beta \neq \beta^\sigma$, then either (a) or (b) of Definition 2 is violated. It is also apparent from the construction that the map $\sigma \mapsto \beta^\sigma$ is continuous, because $\sigma \mapsto \alpha^\sigma$ is obviously continuous, and the Dirac-measure map $\alpha_{-i} \mapsto \delta_{\alpha_{-i}}$ is continuous. □

Lemma 3. *The set of consistent assessments is compact.*

Proof. Lemma 1 implies that $\times_{i\in I}(\Sigma_i \times \Delta_i^2)$ is a compact metrizable space that contains the set of consistent assessments. Therefore, it is enough to show that the latter is closed. Let $\left(\sigma^n, \beta^n\right)_{n\in\mathbb{N}}$ be a converging sequence of consistent assessments with limit $(\sigma^\infty, \beta^\infty)$. For each $i \in I$, let α_i^n be the first-order belief derived from β_i^n ($n \in \mathbb{N} \cup \{\infty\}$), that is,

$$\alpha_i^n(Y|h) = \beta_i^n(Y \times \Delta_{-i}^1|h)$$

for all $h \in H$, $Y \subseteq Z(h)$. By consistency, for all $n \in \mathbb{N}$, $i \in I$, $h \in H$, $a \in A(h)$, we have:

- (a.n) $\alpha_i^n(a|h) = \beta_i^n\left(Z(h,a) \times \Delta_{-i}^1|h\right) = \prod_{j\in I}\sigma_j^n(a_j|h)$,
- (b.n) $\operatorname{marg}_{\Delta_{-i}^1}\beta_i^n(\cdot|h) = \delta_{\alpha_{-i}^n}$, where each α_j^n is determined as in (a.n).

Then, for all $i \in I$, $h \in H$, $a \in A(h)$,

$$\alpha_i^\infty(a|h) = \beta_i^\infty\left(Z(h,a) \times \Delta_{-i}^1|h\right) = \prod_{j\in I}\sigma_j^\infty(a_j|h).$$

Furthermore, $\operatorname{marg}_{\Delta_{-i}^1}\beta_i^\infty(\cdot|h) = \delta_{\alpha_{-i}^\infty}$ for all $i \in I$ and $h \in H$, because $\alpha_{-i}^n \to \alpha_{-i}^\infty$ and the marginalization and Dirac maps $\beta_i \mapsto \operatorname{marg}_{\Delta_{-i}^1}\beta_i$ and $\alpha_{-i} \mapsto \delta_{\alpha_{-i}}$ are continuous. □

A.2. Proof of Proposition 1

Fix $i \in I$ arbitrarily. First-order belief α_i is derived from β_i and, by consistency, gives the behavioral strategy profile σ. Therefore, by assumption each $h' \preceq h$ has probability one under α_i, which implies that $\mathbb{E}[\pi_i|h';\alpha_i] = \mathbb{E}[\pi_i;\alpha_i]$, hence $\mathrm{F}_i(h';\alpha_i) = 0$. Since blame is capped by frustration, $u_i(h', a_i';\beta_i) = \mathbb{E}[\pi_i|h';\alpha_i]$. Therefore, sequential rationality of the equilibrium assessment implies that $\operatorname{Supp}\sigma_i(\cdot|h') \subseteq \arg\max_{a_i'\in A_i(h')}\mathbb{E}[\pi_i|\left(h', a_i'\right);\alpha_i]$. If there is randomization only in the last stage (or none at all), then players maximize locally their expected material payoff on the equilibrium path. Hence, the second claim follows by inspection of the definitions of agent form of the material-payoff game and Nash equilibrium. □

A.3. Proof of Proposition 2

Let $(\bar\sigma, \bar\beta) = \left(\bar\sigma_i, \bar\beta_i\right)_{i\in I}$ be the SE of the material-payoff game, which must be in pure strategies (Remark 3). Fix decision utility functions $u_i(h, a_i;\cdot)$ of the ABI, or ABB kind, and a sequence of real numbers $(\varepsilon_n)_{n\in\mathbb{N}}$, with $\varepsilon_n \to 0$ and $0 < \varepsilon_n < \frac{1}{\max_{i\in I, h\in H}|A_i(h)|}$ for all $n \in \mathbb{N}$. Consider the constrained psychological game where players can choose mixed actions in

$$\Sigma_i^n(h) = \{\sigma_i(\cdot|h) \in \Delta(A_i(h)) : \|\sigma_i(\cdot|h) - \bar\sigma_i(\cdot|h)\| \leq \varepsilon_n\}$$

if h is on the $\bar\sigma$-path, and in

$$\Sigma_i^n(h) = \{\sigma_i(\cdot|h) \in \Delta(A_i(h)) : \forall a_i \in A_i(h), \sigma_i(a_i|h) \geq \varepsilon_n\}$$

if h is off the $\bar\sigma$-path. By construction, these sets are non-empty, convex, and compact. Since the decision utility functions are continuous, and the consistent assessment map $\sigma \longmapsto \beta^\sigma$ is continuous (Lemma 2), the correspondence

$$\sigma \longmapsto \times_{h\in H}\times_{i\in I}\arg\max_{\sigma_i'(\cdot|h)\in\Sigma_i^n(h)}\sum_{a_i\in A_i(h)}\sigma_i'(a_i|h)u_i(h, a_i;\beta_i^\sigma)$$

is upper-hemicontinuous, non-empty, convex, and compact valued; therefore (by Kakutani's theorem), it has a fixed point σ^n. By Lemma 3, the sequence of consistent assessments $\left(\sigma^n, \beta^{\sigma^n}\right)_{n=1}^{\infty}$ has a limit point (σ^*, β^*), which is consistent too. By construction, $\bar{\sigma}(\cdot|h) = \sigma^*(\cdot|h)$ for h on the $\bar{\sigma}$-path, therefore $(\bar{\sigma}, \bar{\beta})$ and (σ^*, β^*) are realization-equivalent. We let $\bar{\alpha}_i$ (respectively, α_i^*) denote the first-order beliefs of i implied by $(\bar{\sigma}, \bar{\beta})$ (respectively, (σ^*, β^*)).

We claim that the consistent assessment (σ^*, β^*) is an SE of the psychological game with decision utility functions $u_i(h, a_i; \cdot)$. We must show that (σ^*, β^*) satisfies sequential rationality. If h is off the $\bar{\sigma}$-path, sequential rationality is satisfied by construction. Since $\bar{\sigma}$ is deterministic and there are no chance moves, if h is on the $\bar{\sigma}$-path (*i.e.*, on the σ^*-path) it must have unconditional probability 1 according to each player's beliefs and there cannot be any frustration; hence, $u_i(h, a_i; \beta_i^*) = \mathbb{E}[\pi_i|h, a_i; \alpha_i^*]$ $(i \in I)$ where α_i^* is determined by σ^*. If, furthermore, it is the second stage ($h = \bar{a}^1$, with $\bar{\sigma}(\bar{a}^1|\varnothing) = 1$), then —by construction— $\mathbb{E}[\pi_i|h, a_i; \alpha_i^*] = \mathbb{E}[\pi_i|h, a_i; \bar{\alpha}_i]$, where $\bar{\alpha}_i$ is determined by $\bar{\sigma}$. Since $\bar{\sigma}$ is an SE of the material-payoff game, sequential rationality is satisfied at h. Finally, we claim that (σ^*, β^*) satisfies sequential rationality also at the root $h = \varnothing$. Let $\iota(h)$ denote the active player at h. Since $\iota(\varnothing)$ cannot be frustrated at $\varnothing$, we must show that action $\bar{a}^1$ with $\bar{\sigma}(\bar{a}^1|\varnothing) = 1$ maximizes his expected material payoff given belief $\alpha^*_{\iota(\varnothing)}$. According to ABB and ABI, player $\iota(a^1)$ can only blame the first mover $\iota(\varnothing)$ and possibly hurt him, if he is frustrated. Therefore, in assessment (σ^*, β^*) at node a^1, either $\iota(a^1)$ plans to choose his (unique) payoff maximizing action, or he blames $\iota(\varnothing)$ strongly enough to give up some material payoff in order to bring down the payoff of $\iota(\varnothing)$. Hence, $\mathbb{E}[\pi_{\iota(\varnothing)}|a^1; \alpha^*_{\iota(a^1)}] \leq \mathbb{E}[\pi_{\iota(\varnothing)}|a^1; \bar{\alpha}_{\iota(a^1)}]$ (anger). By consistency of (σ^*, β^*) and $(\bar{\sigma}, \bar{\beta})$, $\alpha^*_{\iota(a^1)} = \alpha^*_{\iota(\varnothing)}$ and $\bar{\alpha}_{\iota(a^1)} = \bar{\alpha}_{\iota(\varnothing)}$ (cons.). Since (σ^*, β^*) is realization-equivalent (r.e.) to $(\bar{\sigma}, \bar{\beta})$, which is the material-payoff equilibrium (m.eq.), for each $a^1 \in A(\varnothing)$,

$$\mathbb{E}[\pi_{\iota(\varnothing)}|\bar{a}^1; \alpha^*_{\iota(\varnothing)}] \overset{\text{(r.e.)}}{=} \mathbb{E}[\pi_{\iota(\varnothing)}|\bar{a}^1; \bar{\alpha}_{\iota(\varnothing)}] \overset{\text{(m.eq.)}}{\geq}$$
$$\mathbb{E}[\pi_{\iota(\varnothing)}|a^1; \bar{\alpha}_{\iota(\varnothing)}] \overset{\text{(cons.)}}{=} \mathbb{E}[\pi_{\iota(\varnothing)}|a^1; \bar{\alpha}_{\iota(a^1)}] \overset{\text{(anger)}}{\geq}$$
$$\mathbb{E}[\pi_{\iota(\varnothing)}|a^1; \alpha^*_{\iota(a^1)}] \overset{\text{(cons.)}}{=} \mathbb{E}[\pi_{\iota(\varnothing)}|a^1; \alpha^*_{\iota(\varnothing)}].$$

This completes the proof for ABB & ABI. If there are only two players, then we have a LF-game and SA is equivalent to ABB, so (σ^*, β^*) is an SE in this case too. □

A.4. Proof of Proposition 3

Recall that ℓ and f respectively denote the leader and the follower and that, by convention, the leader has no terminating action. By Remark 3, we obtain the unique and pure material-payoff SE strategy pair, viz. $\left(\bar{\sigma}_\ell, \bar{\sigma}_f\right)$, by backward induction: for each $a_\ell \in A_\ell(\varnothing)$, let

$$\bar{s}_f(a_\ell) = \arg \max_{a_f \in A_f(a_\ell)} \pi_f\left(a_\ell, a_f\right)$$

denote the material best reply of f to a_ℓ (unique by assumption); then

$$\forall a_\ell \in A_\ell(\varnothing), \bar{\sigma}_f\left(\bar{s}_f(a_\ell)|a_\ell\right) = 1,$$
$$\bar{\sigma}_\ell\left(\arg \max_{a_\ell \in A_\ell(\varnothing)} \pi_\ell\left(a_\ell, \bar{s}_f(a_\ell)\right)|\varnothing\right) = 1. \tag{11}$$

Let $\left(\bar{a}_\ell, \bar{s}_f\right)$ denote this pure-strategy equilibrium. We must prove that $\mathbb{E}\left[\pi_f; \beta\right] \geq \pi_f\left(\bar{a}_\ell, \bar{s}_f(\bar{a}_\ell)\right)$ for every SE assessment (σ, β). Now, fix arbitrarily an SE (σ, β), thus, by consistency, σ is derived from the first-order beliefs implied by β. Observe that

$$\forall a_\ell \in A_\ell(\varnothing) \setminus \{\bar{a}_\ell\},\ \mathbb{E}\left[\pi_\ell|a_\ell; \beta\right] \overset{\text{(anger)}}{\leq} \pi_\ell\left(a_\ell, \bar{s}_f(a_\ell)\right) \overset{\text{(m.eq.)}}{<} \pi_\ell\left(\bar{a}_\ell, \bar{s}_f(\bar{a}_\ell)\right). \tag{12}$$

The first inequality holds because either f—upon observing a_ℓ—is angry enough to deviate from his material-payoff maximizing action $\bar{s}_f(a_\ell)$ and punish ℓ, or he replies with $\bar{s}_f(a_\ell)$; the second inequality holds because $\left(\bar{a}_\ell, \bar{s}_f\right)$ is the unique material-payoff SE. Since in every SE the leader, who cannot be frustrated, maximizes his expected material payoff, it must be the case that either (i) $\sigma_\ell(\bar{a}_\ell|\varnothing) = 1$, or (ii) $\mathbb{E}\left[\pi_\ell|\bar{a}_\ell; \beta\right] < \pi_\ell\left(\bar{a}_\ell, \bar{s}_f(\bar{a}_\ell)\right)$, otherwise (12) implies that he would choose $\bar{a}_\ell$ with probability 1. In case (i), f cannot be frustrated after $\bar{a}_\ell$; hence, $\sigma_f\left(\bar{s}_f(\bar{a}_\ell)|\bar{a}_\ell\right) = 1$, $\sigma_\ell(\bar{a}_\ell|\varnothing) = 1$, and the SE payoff of f is $\pi_f\left(\bar{a}_\ell, \bar{s}_f(\bar{a}_\ell)\right)$, the same as in the material-payoff SE. In case (ii), $\sigma_f\left(\bar{s}_f(\bar{a}_\ell)|\bar{a}_\ell\right) < 1$, that is, f is not choosing his material-payoff maximizing action $\bar{s}_f(\bar{a}_\ell)$ because he is frustrated. Therefore,

$$\mathbb{E}\left[\pi_f; \beta\right] - \pi_f\left(\bar{a}_\ell, \bar{s}_f(\bar{a}_\ell)\right) = \mathbb{E}\left[\pi_f; \beta\right] - \max_{a_f \in A_f(\bar{a}_\ell)} \pi_f\left(\bar{a}_\ell, a_f\right) = \mathrm{F}_f(\bar{a}_\ell; \alpha) > 0,$$

where α is derived from β. Thus, in each case $\mathbb{E}\left[\pi_f; \beta\right] \geq \pi_f\left(\bar{a}_\ell, \bar{s}_f(\bar{a}_\ell)\right)$. □

A.5. Proof of Proposition 4

Recall that we are assuming that there are no relevant ties (NRT) and that $(\bar{a}_\ell, \bar{s}_f)$ is the unique and necessarily pure SE strategy pair of the material-payoff game. Thus, $\bar{s}_f$ is the best reply function of the follower, and $\bar{a}_\ell = r_\ell(\bar{s}_\ell)$, where $r_\ell : S_f \to A_\ell(\varnothing)$ is the strategic-form best reply function of the leader. We are going through some steps to construct an assessment $(s^*, \beta^*) = \left(a^*_\ell, s^*_f, \beta^*\right)$ such that s^*_f is a *threat* and (s^*, β^*) is an *SE for every θ_f above a threshold*. This yields part 1 of Proposition 4. Parts 2 and 3 follow from Definition 4 of threat, which implies that $a^*_\ell \neq \bar{a}_\ell$ and $\pi_f(a^*_\ell, s^*_f(a^*_\ell)) = \pi_f(a^*_\ell, \bar{s}_f(a^*_\ell)) > \pi_f(\bar{a}_\ell, \bar{s}_f(\bar{a}_\ell))$.

Let TE be the subset of pure strategy Nash equilibria of the material payoff game (mNEs, for short) such that the follower's strategy is a threat. Since the game has some threat $\hat{s}_f$ (by hypothesis) and $(r_\ell(\hat{s}_f), \hat{s}_f)$ is a mNE, TE is nonempty (Remark 5.*i*). With this, let

$$TE^{**} := \arg \max_{(a_\ell, s_f) \in TE} \pi_f\left(a_\ell, s_f(a_\ell)\right)$$
$$TE^{*} := \arg \max_{(a_\ell, s_f) \in TE^{**}} \pi_\ell\left(a_\ell, s_f(a_\ell)\right).$$

In words, TE^* selects the elements of TE that, lexicographically, first maximize the follower's payoff, and then maximize the leader's payoff.[30] Since TE is finite, TE^{**} is finite and nonempty; thus, $TE^* \neq \varnothing$.

Claim. *Set TE^* has the following property:*

$$\begin{aligned} &\forall (a_\ell, \hat{s}_f) \in TE^* : \forall a'_\ell \in A_\ell(\varnothing) \setminus \{a_\ell\}, \\ &\pi_f(a'_\ell, \bar{s}_f(a'_\ell)) \geq \pi_f(a_\ell, \hat{s}_f(a_\ell)) \Longrightarrow \pi_\ell(a'_\ell, \bar{s}_f(a'_\ell)) \leq \pi_\ell(a_\ell, \hat{s}_f(a_\ell)). \end{aligned} \tag{13}$$

Proof. Proof of the claim We prove the claim by contraposition, that is, we prove that if $(a_\ell, \hat{s}_f)$ violates (13) then $(a_\ell, \hat{s}_f) \notin TE^*$. Since $TE^* \subseteq TE^{**}$, if $(a_\ell, \hat{s}_f) \notin TE^{**}$ then $(a_\ell, \hat{s}_f) \notin TE^*$. Next suppose that $(a_\ell, \hat{s}_f) \in TE^{**}$ and (13) does not hold, i.e., for some $\hat{a}_\ell \in A_\ell(\varnothing) \setminus \{a_\ell\}$,

$$\pi_f(\hat{a}_\ell, \bar{s}_f(\hat{a}_\ell)) \geq \pi_f(a_\ell, \hat{s}_f(a_\ell)) \text{ and } \pi_\ell(\hat{a}_\ell, \bar{s}_f(\hat{a}_\ell)) > \pi_\ell(a_\ell, \hat{s}_f(a_\ell)). \tag{14}$$

Let $\hat{s}^*_f(\hat{a}_\ell) = \bar{s}_f(\hat{a}_\ell)$ and $\hat{s}^*_f(a'_\ell) = \hat{s}_f(a'_\ell)$ for every $a'_\ell \neq \hat{a}_\ell$ (thus, including $a'_\ell = a_\ell$). We are going to prove that $(\hat{a}_\ell, \hat{s}^*_f) \in TE^{**}$ as well. This implies that $(a_\ell, \hat{s}_f) \notin TE^*$, because—by $\hat{s}^*_f(\hat{a}_\ell) = \bar{s}_f(\hat{a}_\ell)$ and (14)—the leader's payoff is higher under $(\hat{a}_\ell, \hat{s}^*_f)$ than under $(a_\ell, \hat{s}_f)$. Indeed, $(\hat{a}_\ell, \hat{s}^*_f) \in TE^{**}$ because

$$\pi_f(\hat{a}_\ell, \hat{s}^*_f(\hat{a}_\ell)) = \pi_f(\hat{a}_\ell, \bar{s}_f(\hat{a}_\ell)) \geq \pi_f(a_\ell, \hat{s}_f(a_\ell)) = \pi_f(a_\ell, \bar{s}_f(a_\ell)) > \pi_f(\bar{a}_\ell, \bar{s}_f(\bar{a}_\ell)),$$

where the first equality and inequality hold by $\hat{s}^*_f(\hat{a}_\ell) = \bar{s}_f(\hat{a}_\ell)$ and (14), and the second equality and inequality hold because, since $(a_\ell, \hat{s}_f) \in TE^{**}$, $\hat{s}_f$ is a threat. Thus, $(\hat{a}_\ell, \hat{s}^*_f)$ satisfies Condition 2 of Definition 4 and—by Remark 5.*ii*—strategy $\hat{s}^*_f$ is a threat. Since $(\hat{a}_\ell, \hat{s}^*_f)$ gives to the follower a payoff as high as $(a_\ell, \hat{s}_f) \in TE^{**}$, it follows that $(\hat{a}_\ell, \hat{s}^*_f) \in TE^{**}$. This concludes the proof of the Claim. □

Fix any $(a^*_\ell, \hat{s}_f) \in TE^*$. Note that, if the follower expects to get $\pi_f(a^*_\ell, \hat{s}_f(a^*_\ell))$, then he is frustrated by an action a_ℓ if and only if $\pi_f(a_\ell, \bar{s}_f(a_\ell)) < \pi_f(a^*_\ell, \hat{s}_f(a^*_\ell))$, i.e., his best payoff after a_ℓ is less than he expected. Now consider the off-path modification s^*_f of $\hat{s}_f$ that *minimizes the leaders' payoff in the "cheapest" way* if he deviates from a^*_ℓ in a way that frustrates the follower, and selects the material best reply otherwise. Formally, for every $a_\ell \in A_\ell(\varnothing)$ let

$$A_f^{\min \pi_\ell}(a_\ell) := \arg \min_{a_f \in A_f(a_\ell)} \pi_\ell\left(a_\ell, a_f\right)$$

denote the set of follower's actions that minimize the leader's payoff given a_ℓ. With this, let

$$\left(\pi_f\left(a_\ell, \bar{s}_f(a_\ell)\right) < \pi_f\left(a^*_\ell, \hat{s}_f(a^*_\ell)\right)\right) \Rightarrow s^*_f(a_\ell) = \arg \max_{a_f \in A_f^{\min \pi_\ell}(a_\ell)} \pi_f\left(a_\ell, a_f\right) \tag{15}$$

[30] Note that NRT allows for ties between payoffs of the leader following distinct replies to the same action a_ℓ. Hence, there may be room for such maximization.

$$\left(\pi_f\left(a_\ell, \bar{s}_f\left(a_\ell\right)\right) \geq \pi_f\left(a_\ell^*, \hat{s}_f\left(a_\ell^*\right)\right)\right) \Rightarrow s_f^*\left(a_\ell\right) = \bar{s}_f\left(a_\ell\right) \tag{16}$$

for every $a_\ell \in A_\ell(\varnothing)$. Since $\left(a_\ell^*, \hat{s}_f\right)$ is an mNE, the follower chooses the material best reply on path: $\hat{s}_f\left(a_\ell^*\right) = \bar{s}_f\left(a_\ell^*\right)$. Therefore, (16) implies $s_f^*\left(a_\ell^*\right) = \hat{s}_f\left(a_\ell^*\right)$.

Claim. *Strategy s_f^* threat.*

Proof. Proof of the claim Note that

$$\pi_f\left(a_\ell^*, s_f^*\left(a_\ell^*\right)\right) = \pi_f\left(a_\ell^*, \hat{s}_f\left(a_\ell^*\right)\right) = \pi_f\left(a_\ell^*, \bar{s}_f\left(a_\ell^*\right)\right) > \pi_f(\bar{a}_\ell, \bar{s}_f(\bar{a}_\ell)),$$

by construction and because $\hat{s}_f$ is a threat. Hence, s_f^* satisfies Condition 2 of Definition 4 and—by Remark 5.*ii*—strategy s_f^* is a threat as well. □

Claim. *Action a_ℓ^* is the best reply (unique by NRT) to strategy s_f^*, that is, $a_\ell^* = r_\ell\left(s_f^*\right)$.*

Proof. Proof of the claim For every $a_\ell \in A_\ell(\varnothing)$,

$$\pi_\ell\left(a_\ell^*, s_f^*\left(a_\ell^*\right)\right) = \pi_\ell\left(a_\ell^*, \hat{s}_f\left(a_\ell^*\right)\right) \geq \pi_\ell\left(a_\ell, \hat{s}_f\left(a_\ell\right)\right) \geq \pi_\ell\left(a_\ell, s_f^*\left(a_\ell\right)\right),$$

where the first equality holds by construction, the first inequality holds because $\left(a_\ell^*, \hat{s}_f\right)$ is an mNE, and the second inequality holds by construction. To see the latter note that if the follower is frustrated, then the leader is minimized by s_f^*, and if he is not frustrated, then $s_f^*\left(a_\ell\right) = \bar{s}_f\left(a_\ell\right)$ and the inequality is implied by (13). Thus, a_ℓ^* is the (necessarily unique) best reply to s_f^*. □

Let α_f^* denote the first-order belief of the follower that corresponds to $\left(a_\ell^*, s_f^*\right)$, so that $\alpha_f^*\left(a_\ell^* | \varnothing\right) = 1$ and $\alpha_f^*\left(s_f^*\left(a_\ell\right) | a_\ell\right) = 1$ for each $a_\ell \in A_\ell(\varnothing)$.

Claim. *There is $\underline{\theta}_f \geq 0$ such that, for every $\theta_f \geq \underline{\theta}_f$ strategy s_f^* is a sequential best reply under SA, that is,*

$$\forall a_\ell \in A_\ell(\varnothing),\ s_f^*\left(a_\ell\right) \in \arg\max_{a_f \in A_f\left(a_\ell\right)} \left[\pi_f\left(a_\ell, a_f\right) - \theta_f \mathrm{F}_f\left(a_\ell; \alpha_f^*\right) \pi_\ell\left(a_\ell, a_f\right)\right]. \tag{17}$$

Proof. Proof of the claim Eq. (17) holds by construction for every $\theta_f \geq 0$ at each a_ℓ such that $\pi_f\left(a_\ell, \bar{s}_f\left(a_\ell\right)\right) \geq \pi_f\left(a_\ell^*, s_f^*\left(a_\ell^*\right)\right)$, because $\mathrm{F}_f\left(a_\ell; \alpha_f^*\right) = 0$ and $s_f^*\left(a_\ell\right) = \bar{s}_f\left(a_\ell\right)$ is the material best reply to a_ℓ. Next consider any a_ℓ such that $\pi_f\left(a_\ell, \bar{s}_f\left(a_\ell\right)\right) < \pi_f\left(a_\ell^*, s_f^*\left(a_\ell^*\right)\right)$. Then,

$$\begin{aligned}\mathrm{F}_f\left(a_\ell; \alpha_f^*\right) &= \pi_f\left(a_\ell^*, s_f^*\left(a_\ell^*\right)\right) - \max_{a_f \in A_f\left(a_\ell\right)} \pi_f\left(a_\ell, a_f\right) \\ &= \pi_f\left(a_\ell^*, s_f^*\left(a_\ell^*\right)\right) - \pi_f\left(a_\ell, \bar{s}_f\left(a_f\right)\right) > 0.\end{aligned}$$

For every such action a_ℓ, we want to show that there is a large enough $\theta_f \geq 0$ so that, for every $a_f \in A_f\left(a_\ell\right) \setminus \left\{s_f^*\left(a_\ell\right)\right\}$,

$$\pi_f\left(a_\ell, s_f^*\left(a_\ell\right)\right) - \theta_f \mathrm{F}_f\left(a_\ell; \alpha_f^*\right) \pi_\ell\left(a_\ell, s_f^*\left(a_\ell\right)\right) \geq \pi_f\left(a_\ell, a_f\right) - \theta_f \mathrm{F}_f\left(a_\ell; \alpha_f^*\right) \pi_\ell\left(a_\ell, a_f\right)$$

that is

$$\theta_f \mathrm{F}_f\left(a_\ell; \alpha_f^*\right) \left[\pi_\ell\left(a_\ell, a_f\right) - \pi_\ell\left(a_\ell, s_f^*\left(a_\ell\right)\right)\right] \geq \pi_f\left(a_\ell, a_f\right) - \pi_f\left(a_\ell, s_f^*\left(a_\ell\right)\right). \tag{18}$$

Since $s_f^*\left(a_\ell\right) \in A_f^{\min \pi_\ell}\left(a_\ell\right)$, $\pi_\ell\left(a_\ell, a_f\right) - \pi_\ell\left(a_\ell, s_f^*\left(a_\ell\right)\right) \geq 0$. Therefore

$$\theta_f \mathrm{F}_f\left(a_\ell; \alpha_f^*\right) \left[\pi_\ell\left(a_\ell, a_f\right) - \pi_\ell\left(a_\ell, s_f^*\left(a_\ell\right)\right)\right] \geq 0.$$

If $\pi_\ell\left(a_\ell, a_f\right) - \pi_\ell\left(a_\ell, s_f^*\left(a_\ell\right)\right) = 0$, then also $a_f \in A_f^{\min \pi_\ell}\left(a_\ell\right)$ is a minimizer like $s_f^*\left(a_\ell\right)$. Since $s_f^*\left(a_\ell\right) \in \arg\max_{a_f \in A_f^{\min \pi_\ell}\left(a_\ell\right)} \pi_f\left(a_\ell, a_f\right)$, we have $0 \geq \pi_f\left(a_\ell, a_f\right) - \pi_f\left(a_\ell, s_f^*\left(a_\ell\right)\right)$ and (18) must hold. If $\pi_\ell\left(a_\ell, a_f\right) - \pi_\ell\left(a_\ell, s_f^*\left(a_\ell\right)\right) > 0$, since $\mathrm{F}_f\left(a_\ell; \alpha_f^*\right) > 0$ as well, we obtain the pairwise threshold

$$\underline{\theta}_f\left(a_\ell, a_f\right) := \max\left\{0, \frac{\pi_f\left(a_\ell, a_f\right) - \pi_f\left(a_\ell, s_f^*\left(a_\ell\right)\right)}{\mathrm{F}_f\left(a_\ell; \alpha_f^*\right)\left[\pi_\ell\left(a_\ell, a_f\right) - \pi_\ell\left(a_\ell, s_f^*\left(a_\ell\right)\right)\right]}\right\}.$$

With this, (17) holds if

$$\theta_f \geq \underline{\theta}_f := \max\left\{\underline{\theta}_f\left(a_\ell, a_f\right) : \pi_f\left(a_\ell, \bar{s}_f\left(a_\ell\right)\right) < \pi_f\left(a_\ell^*, s_f^*\left(a_\ell^*\right)\right), a_f \in A_f\left(a_\ell\right) \setminus A_f^{\min \pi_\ell}\left(a_\ell\right)\right\} \geq 0. \quad \square$$

Let $s^* = \left(a^*, s_f^*\right)$ and let $\beta^* = \beta^{s^*}$ be the unique second-order belief system that makes assessment (s^*, β^*) consistent (see Lemma 2). If $\theta_f \geq \underline{\theta}_f$, the last two claims imply that (s^*, β^*) is an SE under SA, which assumes $\mathrm{F}_f = \mathrm{B}_f$. By Remark 4, in LF-games frustration and blame coincide—hence (s^*, β^*) is an SE— under ABB as well. To sum up, we proved that, for every $\theta_f \geq \underline{\theta}_f$, (s^*, β^*) is an SE under ABB/SA where s_f^* is a threat. $\square$

A.6. Proof of Proposition 5

By Proposition 4, in the given LF-Game with a threat and with no relevant ties, there is a pure strategy deterrence SE with a threat $\left(\hat{s}, \hat{\beta}\right)$ for a sufficiently high θ_f. Thus, the set of parameters that support $\left(\hat{s}, \hat{\beta}\right)$ as an SE of the game with anger is nonempty. The threshold we are analyzing is

$$\underline{\theta}_f := \inf\left\{\theta_f \in \mathbb{R}_+ : \left(\hat{s}, \hat{\beta}\right) \text{ is SE given } \theta_f\right\},$$

which, of course, depends on the material payoffs. We derive an explicit expression for $\underline{\theta}_f$.

Let $\hat{\alpha}$ denote the corresponding profile of first-order beliefs. As in the proof of Proposition 4 recall that $\left(\bar{a}_\ell, \bar{s}_f\right)$ denotes the (necessarily pure) material-payoff SE strategy pair. Furthermore, $\hat{s}_f$ is a threat. Suppose that the leader deviates to the material-payoff SE action $\bar{a}_\ell$. The follower's maximal material payoff after $\bar{a}_\ell$ is $\pi_f(\bar{a}_\ell, \bar{s}_f(\bar{a}_\ell))$, while her payoff from the deterrence SE strategy $\hat{s}_f$ in response to $\bar{a}_\ell$ is $\pi_f(\bar{a}_\ell, \hat{s}_f(\bar{a}_\ell))$. The follower's frustration after $\bar{a}_\ell$ is therefore

$$\mathrm{F}_f\left(\bar{a}_\ell; \hat{\alpha}_f\right) = \pi_f(\hat{a}_\ell, \hat{s}_f\left(\hat{a}_\ell\right)) - \pi_f(\bar{a}_\ell, \bar{s}_f\left(\bar{a}_\ell\right)) > 0. \tag{19}$$

Since the threat $\hat{s}_f$ of the follower is sequentially rational (given θ_f) after the deviation to $\bar{a}_\ell$, it must be the case that

$$u_f(\bar{a}_\ell, \hat{s}_f(\bar{a}_\ell); \hat{\alpha}_f) \geq u_f(\bar{a}_\ell, \bar{s}_f(\bar{a}_\ell); \hat{\alpha}_f)$$

and therefore that

$$\pi_f(\bar{a}_\ell, \hat{s}_f(\bar{a}_\ell)) - \theta_f \mathrm{F}_f\left(\bar{a}_\ell; \hat{\alpha}_f\right)\pi_\ell(\bar{a}_\ell, \hat{s}_f(\bar{a}_\ell)) \geq \pi_f(\bar{a}_\ell, \bar{s}_f(\bar{a}_\ell)) - \theta_f \mathrm{F}_f\left(\bar{a}_\ell; \hat{\alpha}_f\right)\pi_\ell(\bar{a}_\ell, \bar{s}_f(\bar{a}_\ell))$$

rearranging this expression and collecting terms we have

$$\theta_f \mathrm{F}_f\left(\bar{a}_\ell; \hat{\alpha}_f\right)\left(\pi_\ell(\bar{a}_\ell, \bar{s}_f(\bar{a}_\ell)) - \pi_\ell(\bar{a}_\ell, \hat{s}_f(\bar{a}_\ell))\right) \geq \pi_f(\bar{a}_\ell, \bar{s}_f(\bar{a}_\ell)) - \pi_f(\bar{a}_\ell, \hat{s}_f(\bar{a}_\ell))$$

The right-hand side of this expression is positive because $\bar{s}_f$ is the material-payoff SE strategy of the follower and hence maximizes $\pi_f(\cdot)$ after $\bar{a}_\ell$. This implies that $\pi_\ell(\bar{a}_\ell, \bar{s}_f(\bar{a}_\ell)) - \pi_\ell(\bar{a}_\ell, \hat{s}_f(\bar{a}_\ell))$ is also positive, since $\theta_f > 0$ and $\mathrm{F}_f\left(\bar{a}_\ell; \hat{\alpha}_f\right) > 0$. Moving all terms besides θ_f to the righthand side we have

$$\theta_f \geq \frac{\pi_f(\bar{a}_\ell, \bar{s}_f(\bar{a}_\ell)) - \pi_f(\bar{a}_\ell, \hat{s}_f(\bar{a}_\ell))}{\mathrm{F}_f\left(\bar{a}_\ell; \hat{\alpha}_f\right)\left(\pi_\ell(\bar{a}_\ell, \bar{s}_f(\bar{a}_\ell)) - \pi_\ell(\bar{a}_\ell, \hat{s}_f(\bar{a}_\ell))\right)}$$

and substituting in the follower's frustration after the leader's deviation to the material-payoff SE action $\bar{a}_\ell$ we have

$$\theta_f \geq \frac{\pi_f(\bar{a}_\ell, \bar{s}_f(\bar{a}_\ell)) - \pi_f(\bar{a}_\ell, \hat{s}_f(\bar{a}_\ell))}{\left(\pi_f(\hat{a}_\ell, \hat{s}_f\left(\hat{a}_\ell\right)) - \pi_f(\bar{a}_\ell, \bar{s}_f\left(\bar{a}_\ell\right))\right)\left(\pi_\ell(\bar{a}_\ell, \bar{s}_f(\bar{a}_\ell)) - \pi_\ell(\bar{a}_\ell, \hat{s}_f(\bar{a}_\ell))\right)} = \underline{\theta}_f. \tag{20}$$

That is, the inf is a min, by continuity of decision utility in θ_f. With this, the proposition may be verified by inspection.of the expression for $\underline{\theta}_f$. $\square$

A.7. Proof of Proposition 6

Let $\hat{\sigma} = (\hat{a}_\ell, \hat{s}_f)$ be a pure SE strategy pair with SA/ABB or ABI of a LF-game with no relevant ties, and let $\hat{\alpha}$ denote the corresponding profile of first-order beliefs. First note that the consistency condition of SE implies that deviations from $\hat{a}_\ell$ would be rated by f as unintended mistakes, hence f would not be angry under ABI. Therefore, there the pure SE with ABI is unique and it coincides with $\hat{\sigma}$. Next we consider SA/ABB and we prove the result by contraposition, that is, we show that if path $(\hat{a}_\ell, \hat{s}_f(\hat{a}_\ell))$ differs from the material-payoff equilibrium path $(\bar{a}_\ell, \bar{s}_f(\bar{a}_\ell))$, then there is a threat.

As the leader cannot be frustrated at the root, and the follower cannot be frustrated after $\hat{a}_\ell$, which is expected with probability 1, we must have $\hat{a}_\ell = r_\ell(\hat{s}_f)$ and $\hat{s}_f(\hat{a}_\ell) = \arg\max_{a_f \in A(\hat{a}_\ell)} \pi_f(\hat{a}_\ell, a_f) = \bar{s}_f(\hat{a}_\ell)$. Suppose $(\hat{a}_\ell, \hat{s}_f(\hat{a}_\ell)) \neq (\bar{a}_\ell, \bar{s}_f(\bar{a}_\ell))$. Then $\hat{a}_\ell \neq \bar{a}_\ell$ and

$$\pi_\ell(\hat{a}_\ell, \bar{s}_f(\hat{a}_\ell)) = \pi_\ell(\hat{a}_\ell, \hat{s}_f(\hat{a}_\ell)) > \pi_\ell(\bar{a}_\ell, \hat{s}_f(\bar{a}_\ell)), \tag{21}$$

because $\hat{a}_\ell = r_\ell(\hat{s}_f)$. On the other hand, since $(\bar{a}_\ell, \bar{s}_f)$ is the unique material-payoff equilibrium

$$\pi_\ell(\hat{a}_\ell, \bar{s}_f(\hat{a}_\ell)) \overset{\text{(m-eq)}}{<} \pi_\ell(\bar{a}_\ell, \bar{s}_f(\bar{a}_\ell)).$$

Therefore,

$$\hat{s}_f(\bar{a}_\ell) \neq \bar{s}_f(\bar{a}_\ell) = \arg\max_{a_f \in A(\bar{a}_\ell)} \pi_f(\bar{a}_\ell, a_f),$$

that is, the follower is not choosing his material-payoff best response. This can happen in an SE with SA/ABB only if the follower is frustrated after $\bar{a}_f$, that is,

$$\mathrm{F}_f(\bar{a}_\ell; \hat{\alpha}_f) = \pi_f(\hat{a}_\ell, \hat{s}_f(\hat{a}_\ell)) - \max_{a_f \in A_f(\bar{a}_\ell)} \pi_f(\bar{a}_\ell, a_f) = \pi_f(\hat{a}_\ell, \hat{s}_f(\hat{a}_\ell)) - \pi_f(\bar{a}_\ell, \bar{s}_f(\bar{a}_\ell)) > 0.$$

Taking into account that $\hat{s}_f(\hat{a}_\ell) = \bar{s}_f(\hat{a}_\ell)$, this shows that Condition 2 of Definition 4 of threat holds. Furthermore, the material-payoff SE condition for the leader and (21) yield

$$\pi_\ell(\bar{a}_\ell, \bar{s}_f(\bar{a}_\ell)) \overset{\text{(m-eq)}}{\geq} \pi_\ell(\hat{a}_\ell, \bar{s}_f(\hat{a}_\ell)) \overset{(21)}{=} \pi_\ell(\hat{a}_\ell, \hat{s}_f(\hat{a}_\ell)) \overset{(21)}{>} \pi_\ell(\bar{a}_\ell, \hat{s}_f(\bar{a}_\ell)).$$

Therefore Condition 1 holds as well. □

References

Aina, C., Battigalli, P., Gamba, A., 2018. Frustration and Anger in the Ultimatum Game: An Experiment. IGIER working paper 621.
Akerlof, R.J., 2016. Anger and enforcement. J. Econ. Behav. Organ. 126, 110–124.
Alicke, M.D., 2000. Culpable control and the psychology of blame. Psychol. Bull. 126, 556.
Aliprantis, C.R., Border, K.C., 2006. Infinite Dimensional Analysis. Springer, Berlin.
Anderson, E., Simester, D., 2010. Price stickiness and customer antagonism. Q. J. Econ. 125, 729–765.
Averill, J.R., 1982. Anger & Aggression: An Essay on Emotion. Springer, New York.
Bartling, B., Fischbacher, U., 2012. Shifting the blame: on delegation and responsibility. Rev. Econ. Stud. 79, 67–87.
Battigalli, P., 1997. On rationalizability in extensive games. J. Econ. Theory 74, 40–61.
Battigalli, P., Corrao, R., Dufwenberg, M., 2019. Incorporating belief-dependent motivation in games. J. Econ. Behav. Organ.. in press.
Battigalli, P., Dufwenberg, M., 2007. Guilt in games. Am. Econ. Rev. Pap. Proc. 97, 170–176.
Battigalli, P., Dufwenberg, M., 2009. Dynamic psychological games. J. Econ. Theory 144, 1–35.
Battigalli, P., Siniscalchi, M., 1999. Hierarchies of conditional beliefs and interactive epistemology in dynamic games. J. Econ. Theory 88, 188–230.
Baumeister, R.F., Smart, L., Boden, J.M., 1996. Relation of threatened egotism to violence and aggression: the dark side of high self-esteem. Psychol. Rev. 103, 5–33.
Bell, D., 1985. Disappointment in decision making under uncertainty. Oper. Res. 33, 1–27.
Berg, J., Dickhaut, J., McCabe, K., 1995. Trust, reciprocity, and social history. Games Econ. Behav. 10, 122–142.
Berkowitz, L., 1978. Whatever happened to the frustration-aggression hypothesis? Am. Behav. Sci. 21, 691–708.
Berkowitz, L., 1989. Frustration-aggression hypothesis: examination and reformulation. Psychol. Bull. 106, 59–73.
Bolton, G.E., Ockenfels, A., 1989. ERC: a theory of equity, reciprocity, and competition. Am. Econ. Rev. 90, 166–193.
Bosman, R., Sutter, M., van Winden, F., 2005. The impact of real effort and emotions in the power-to-take game. J. Econ. Psychol. 26, 407–429.
Bosman, R., van Winden, F., 2002. Emotional hazard in a power-to-take experiment. Econ. J. 112, 147–169.
Brandts, J., Sola, C., 2001. Reference points and negative reciprocity in simple sequential games. Games Econ. Behav. 36, 138–157.
Buss, D., 2016. Evolutionary Psychology. The New Science of the Mind, 5th edition. Routledge, New York.
Brams, S., 2011. Game Theory and the Humanities: Bridging Two Worlds. MIT Press, Cambridge, MA.
Card, D., Dahl, G., 2011. Family violence and football: the effect of unexpected emotional cues on violent behavior. Q. J. Econ. 126, 103–143.
Carpenter, J., Matthews, P., 2012. Norm enforcement: anger, indignation or reciprocity. J. Eur. Econ. Assoc. 10, 555–572.
Celen, B., Schotter, A., Blanco, M., 2017. On blame and reciprocity: an experimental study. J. Econ. Theory 169, 62–92.
Chang, L., Smith, A., Dufwenberg, M., Sanfey, A., 2011. Triangulating the neural, psychological, and economic bases of guilt aversion. Neuron 70, 560–572.
Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. Q. J. Econ. 117, 817–869.
Dollard, J., Doob, L., Miller, N., Mowrer, O., Sears, R., 1939. Frustration & Aggression. Yale University Press, New Haven, NJ.
Dufwenberg, M., 2002. Marital investment, time consistency and emotions. J. Econ. Behav. Organ. 48, 57–69.
Dufwenberg, M., Kirchsteiger, G., 2004. A theory of sequential reciprocity. Games Econ. Behav. 47, 268–298.

Dufwenberg, M., Li, F., Smith, A., 2018a. Promises & Punishment, typescript.
Dufwenberg, M., Li, F., Smith, A., 2018b. Threats. typescript.
Elster, J., 1998. Emotions and economic theory. J. Econ. Lit. 36, 47–74.
Falk, A., Fischbacher, U., 2006. A theory of reciprocity. Games Econ. Behav. 54, 293–315.
Falk, A., Fehr, E., Fischbacher, U., 2003. On the nature of fair behavior. Econ. Inq. 41, 20–26.
Falk, A., Fehr, E., Fischbacher, U., 2008. Testing theories of fairness—intentions matter. Games Econ. Behav. 62, 287–303.
Fehr, E., Gächter, S., 2002. Altruistic punishment in humans. Nature 415, 137–140.
Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition, and cooperation. Q. J. Econ. 114, 817–868.
Frank, R.H., 1988. Passions Within Reason: The Strategic Role of the Emotions. W.W. Norton & Co., Chicago.
Frijda, N.H., 1993. The place of appraisal in emotion. Cogn. Emot. 7, 357–387.
Gale, J., Binmore, K.G., Samuelson, L., 1995. Learning to be imperfect: the ultimatum game. Games Econ. Behav. 8, 56–90.
Geanakoplos, J., Pearce, D., Stacchetti, E., 1989. Psychological games and sequential rationality. Games Econ. Behav. 1, 60–79.
Gneezy, U., Imas, A., 2014. Materazzi effect and the strategic use of anger in competitive interactions. Proc. Natl. Acad. Sci. 111, 1334–1337.
Güth, W., Kliemt, H., 1998. The indirect evolutionary approach: bridging between rationality and adaptation. Ration. Soc. 10, 377–399.
Güth, W., Schmittberger, R., Schwarze, B., 1982. An experimental-analysis of ultimatum bargaining. J. Econ. Behav. Organ. 3, 367–388.
Gurdal, M., Miller, J., Rustichini, A., 2014. Why blame? J. Polit. Econ. 121, 1205–1246.
Harmon-Jones, E., Sigelman, J., 2001. State anger and prefrontal brain activity: evidence that insult-related relative left-prefrontal activation is associated with experienced anger and aggression. J. Pers. Soc. Psychol. 80, 797–803.
Halpern, J.Y., 2016. Actual Causality. MIT Press, Cambridge.
Hirshleifer, J., 1987. On the emotions as guarantors of threats and promises. In: Dupre, John (Ed.), The Latest on the Best: Essays on Evolution & Optimality. MIT Press, Cambridge, pp. 307–326.
Klein, D.B., O'Flaherty, B., 1993. A game-theoretic rendering of promises and threats. J. Econ. Behav. Organ. 21, 295–314.
Kőszegi, B., Rabin, M., 2006. A model of reference-dependent preferences. Q. J. Econ. 121, 1133–1166.
Kőszegi, B., Rabin, M., 2007. Reference-dependent risk attitudes. Am. Econ. Rev. 97, 1047–1073.
Kreps, D., Wilson, R., 1982. Sequential equilibria. Econometrica 50, 863–894.
van Leeuwen, B., Noussair, C., Offerman, T., Suetens, S., van Veelen, M., van de Ven, J., 2018. Predictably angry – facial cues provide a credible signal of destructive behavior. Manag. Sci. 64, 2973–3468.
Loomes, G., Sugden, R., 1986. Disappointment and dynamic consistency in choice under uncertainty. Rev. Econ. Stud. 53, 271–282.
Marcus-Newhall, A., Pedersen, W.C., Carlson, M., Miller, N., 2000. Displaced aggression is alive and well: a meta-analytic review. J. Pers. Soc. Psychol. 78, 670–689.
Munyo, I., Rossi, M.A., 2013. Frustration, euphoria, and violent crime. J. Econ. Behav. Organ. 89, 136–142.
Nelson, W.R., 2002. Equity or intention: it is the thought that counts. J. Econ. Behav. Organ. 48, 423–430.
Passarelli, F., Tabellini, G., 2017. Emotions and political unrest. J. Polit. Econ. 125, 903–946.
Persson, E., 2018. Testing the impact of frustration and anger when responsibility is low. J. Econ. Behav. Organ. 145, 435–448.
Pillutla, M., Murningham, K., 1996. Unfairness, anger, and spite: emotional rejections of ultimatum offers. Organ. Behav. Hum. Decis. Process. 68, 208–224.
Potegal, M., Spielberger, C., Stemmler, G., 2010. International Handbook of Anger: Constituent and Concomitant Biological, Psychological, and Social Processes. Springer, New York.
Rabin, M., 1993. Incorporating fairness into game theory and economics. Am. Econ. Rev. 83, 1281–1302.
Rauh, M., Seccia, G., 2006. Anxiety and performance: a learning-by-doing model. Int. Econ. Rev. 47, 583–609.
Reuben, E., van Winden, F., 2008. Social ties and coordination on negative reciprocity: the role of affect. J. Public Econ. 92, 34–53.
Rotemberg, J., 2005. Customer anger at price increases, changes in the frequency of price adjustment and monetary policy. J. Monet. Econ. 52, 829–852.
Rotemberg, J., 2008. Minimally acceptable altruism and the ultimatum game. J. Econ. Behav. Organ. 66, 457–476.
Rotemberg, J., 2011. Fair pricing. J. Eur. Econ. Assoc. 9, 952–981.
Sanfey, A., 2009. Expectations and social decision-making: biasing effects of prior knowledge on ultimatum responses. Mind Soc. 8, 93–107.
Schotter, A., Sopher, B., 2007. Advice and behavior in intergenerational ultimatum games: an experimental approach. Games Econ. Behav. 58, 365–393.
Sell, A., Cosmides, L., Tooby, J., 2009. Formidability and the logic of human anger. Proc. Natl. Acad. Sci. 106, 15073–15078.
Selten, R., 1978. The chain store paradox. Theory Decis. 9, 127–159.
Silfver, M., 2007. Coping with guilt and shame: a narrative approach. J. Moral Educ. 36, 169–183.
Smith, A., 2009. Belief-dependent anger in games. typescript.
Sutter, M., 2007. Outcomes versus intentions: on the nature of fair behavior and its development with age. J. Econ. Psychol. 28, 69–78.
Winter, E., 2014. Feeling Smart: Why Our Emotions Are More Rational Than We Think. PublicAffairs, New York.
Winter, E., Garcia-Jurado, I., Mendez Naya, L., 2016. Mental equilibrium and strategic emotions. Manag. Sci. 63, 1302–1317.
Xiang, T., Lohrenz, T., Montague, R., 2013. Computational substrates of norms and their violations during social exchange. J. Neurosci. 33, 1099–1108.