

Indirect Reciprocity with Simple Records

Daniel Clark^a, Drew Fudenberg^{a,1}, and Alexander Wolitzky^a

^aDepartment of Economics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

This manuscript was compiled on March 10, 2020

Indirect reciprocity is a foundational mechanism of human cooperation. Existing models of indirect reciprocity fail to robustly support social cooperation: image scoring models fail to provide robust incentives, while social standing models are not informationally robust. Here we provide a new model of indirect reciprocity based on simple, decentralized records: each individual's record depends on their own past behavior alone, and not on their partners' past behavior or their partners' partners' past behavior. When social dilemmas exhibit a coordination motive (or *strategic complementarity*), tolerant trigger strategies based on simple records can robustly support positive social cooperation and exhibit strong stability properties. In the opposite case of *strategic substitutability*, positive social cooperation cannot be robustly supported. Thus, the strength of short-run coordination motives in social dilemmas determines the prospects for robust long-run cooperation.

Indirect reciprocity | Robust cooperation | Strategic complementarity | Strategic substitutability |

People (and perhaps also other animals) often trust each other to cooperate even when they know they will never meet again. Such indirect reciprocity relies on individuals having some information about how their partners have behaved in the past. Existing models of indirect reciprocity fall into two paradigms. In the image scoring paradigm, each individual carries an *image* that improves when they help others, and (at least some) individuals help only those with good images (1, 2). In the standing paradigm, each individual carries a *standing* that typically improves when they help others with good standing, but not when they help those with bad standing, and individuals with good standing help only other good-standing individuals (3, 4).

Neither of these paradigms provides a robust explanation for social cooperation. In image-scoring models, there is no reason for an individual to only help partners with good images: since the partner's image does not affect one's future payoff, helping some partners and not others is optimal only if one is completely indifferent between helping and not helping. In game-theoretic terms, individuals never have strict incentives to follow image-scoring strategies, and hence such strategies can form at best a weak equilibrium. Closely related to this point, image-scoring equilibria are unstable in several environments (5, 6). Standing models do yield strict, stable equilibria, but they fail to be informationally robust: an individual's standing is a function of not only their own past behavior, but also their past partners' behavior, their partners' partners' behavior, and so on ad infinitum. In the absence of centralized record-keeping or some way of physically marking bad-standing individuals, computing such a function requires information that is likely unavailable in many groups (7).

We develop a new theoretical paradigm for modeling indirect reciprocity that supports positive social cooperation as a strict, stable equilibrium while relying only on simple, *individualistic* information: when two players meet, they observe each

other's records and nothing else, and each individual's record depends only on their own past behavior. (Individualistic information is also called "first-order" (8–10).)

As our model of individual interaction, we use the classic prisoner's dilemma ("PD") with actions *C*, *D* ("Cooperate," "Defect") and a standard payoff normalization, where the gain from unilateral defection, *g*, and the loss from unilateral cooperation, *l*, are both positive and satisfy the condition $g < l + 1$, which means that joint payoffs are maximized by mutual cooperation—see the leftmost matrices in Fig. 1. This canonical game can capture many two-sided interactions, such as business partnerships (11), management of public resources (12, 13), and risk-sharing in developing societies (14), as well as many well-documented animal behaviors (15).

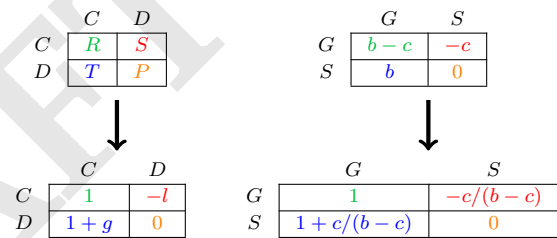


Fig. 1. The prisoner's dilemma. The matrices on the left show how any prisoner's dilemma can be represented by the standard normalization with $g = (T - R)/(R - P)$ and $l = (P - S)/(R - P)$, where $T > R > P > S$. The matrices on the right illustrate this normalization for "donation games" in which choosing *G* (*Give*) instead of *S* (*Shirk*) incurs a personal cost *c* and gives benefit $b > c$ to the opponent.

A critical feature of the PD is whether it exhibits *strategic complementarity* or *strategic substitutability*. Strategic complementarity means that the gain from playing *D* is greater when

Significance Statement

Indirect reciprocity is a foundational mechanism of human cooperation, and understanding the social structures that allow it to arise continues to be a core issue in both the social sciences and evolutionary biology. This paper analyzes a model of indirect reciprocity in steady-state equilibria, where players observe only their partners' records, and each individual's record depends on their own past behavior alone. We show that tolerant trigger strategies based on these simple records can robustly support positive social cooperation in games with sufficient "strategic complementarity," both in the prisoner's dilemma and in some multiplayer public goods games, and we show that the resulting cooperative equilibria have strong stability properties.

All authors contributed equally to the work presented in this paper.

The authors declare no competing financial interests.

¹To whom correspondence should be addressed. E-mail: drew.fudenberg@gmail.com

the opponent also plays D . In the PD payoff matrix displayed in **Fig. 1**, this corresponds to the condition

$$g < l. \quad \text{[Strategic Complementarity]}$$

The opposite case of *strategic substitutability* arises when the gain from playing D is greater when the opponent plays C : mathematically, this occurs when

$$g > l. \quad \text{[Strategic Substitutability]}$$

Many previous studies of indirect reciprocity restrict attention to the “donation game” instance of the PD where $g = l$, as in the rightmost matrices in **Fig. 1** (16).^{*} Our analysis reveals this to be a knife-edge case that obscures the distinction between strategic complementarity ($g < l$) and substitutability ($g > l$). This distinction has long been known to be of critical importance in economics (18, 19), while its implications for cooperation in the repeated prisoner’s dilemma have been noted more recently (8, 9). When a player’s record depends only on their own past actions, the future reward for cooperation (or future penalty for defection) is independent of their current opponent’s record. Therefore, to obtain an equilibrium where a player has a strict incentive to cooperate if and only if the opponent’s record is good, the cost of cooperation must be lower against an opponent with a good record (who cooperates) than against one with a bad record (who defects): that is, cooperation requires $g < l$.

Strategic complementarity is a common case in realistic social dilemmas. It implies that although D is always selfishly optimal (a defining feature of the PD), the social dilemma nonetheless retains some aspect of a coordination game, so that playing C is less costly when one’s partner also plays C . For example, mobbing a predator is always risky (hence costly) for each individual, but it is much less risky when others also mob (20).

In our model, each player’s record is an integer, which evolves as a function of their history of plays of C and D . We assume the system is subject to some noise, so that, whenever an individual plays C , with probability ε their record updates as if they had played D instead.[†] Here the level of noise $\varepsilon \in (0, 1)$ can reflect either errors in recording or errors in executing the intended action.

A simple example of such a record system is the “Counting D ’s” system where a player’s record is just a count of the number of times they have defected (or cooperated and were hit by noise). More complicated record systems could also count the number of times a player cooperated, and could also keep track of the time path of plays of C and D . We will analyze a fairly broad class of strategies, with the following three defining properties: (i) The set of all possible records can be partitioned into two classes, “good records” and “bad records.” (ii) When two players with good records meet each other, they cooperate; if instead either partner has a bad record, both players defect. (iii) The class of bad records is absorbing: once a player obtains a bad record, their record remains bad forever. We refer to this as the class of *trigger strategies*.

Examples of trigger strategies include strategies where a player’s record becomes bad once the absolute number of

times they have defected crosses a threshold K , as well as strategies where their record becomes bad the first time the fraction of times they have defected crosses a threshold. We call strategies of the former type *tolerant grim trigger strategies* or *GrimK*, as they are a form of the well-known grim trigger strategies (21) with a “tolerance” of K recorded plays of D . We will see that *GrimK* strategies succeed in supporting cooperation for a broad range of payoff parameters. Moreover, if the payoff parameters preclude cooperation under *GrimK* strategies, they also preclude cooperation under any other trigger strategy.

We analyze the steady-state equilibria of a system where the total population size is constant, but each individual has a geometrically distributed lifetime with survival probability $\gamma \in (0, 1)$. Players play the PD with random rematching every period, and receive no information about their current partner other than their record. To ensure robustness, we insist that equilibrium behavior is strictly optimal at every record; in classical (normal-form) games, this implies that the equilibrium is evolutionarily stable (22, 23).

Results

Steady-State Cooperation. We show that *GrimK* strategies can form a strict steady-state equilibrium if and only if the PD exhibits substantial strategic complementarity, in that the gain from playing D rather than C is significantly greater when the opponent plays D : the precise condition required in the PD payoff matrix displayed in **Fig. 1** is

$$g < \frac{l}{1+l}.$$

Under this condition, the tolerance level K can be tuned so that *GrimK* strategies support positive social cooperation in a steady-state equilibrium.

To see how to tune the threshold K , note that since even individuals who always try to cooperate are sometimes recorded as playing D due to noise, K must be large enough that the steady-state share of the population with good records is sufficiently high: with any fixed value of K , a population of sufficiently long-lived players would almost all have bad records. However, K also cannot be too high, as otherwise an individual with a very good record (that is, with a very low number of D ’s) can safely play D until their record approaches the threshold. Another constraint is that an individual with record $K - 1$ who meets a partner with a bad record must not be tempted to deviate to C to preserve their own good record. These constraints lead to an upper bound on the maximum share of cooperators in equilibrium. As lifetimes become long and noise becomes small, this upper bound converges to 0 whenever $g > l/(1+l)$, and to $l/(1+l)$ whenever $g < l/(1+l)$ —see Figure 2—and we show that this share of cooperators can in fact be attained in equilibrium in the $(\gamma, \varepsilon) \rightarrow (1, 0)$ limit. Thus, greater strategic complementarity (higher l and lower g) not only helps support some cooperation; it also increases the maximum level of cooperation in the limit, as shown in **Fig. 3**.

We also show that, in the $(\gamma, \varepsilon) \rightarrow (1, 0)$ limit, no trigger strategies can support a positive equilibrium share of cooperators if $g > l/(1+l)$, and no trigger strategies can support an equilibrium share of cooperations greater than $l/(1+l)$ if $g < l/(1+l)$. Thus, when lifetimes are long and noise is small,

^{*}However, the $g \neq l$ case has also received significant attention: for example, the seminal article of Axelrod and Hamilton (17) took $g = 1$ and $l = 1/2$.

[†]It would not substantively affect our results to assume that there is also noise when an individual plays D , so we exclude this possibility for simplicity.

167 *GrimK* strategies attain optimum equilibrium cooperation
 168 within the class of trigger strategies. The logic of this result is
 169 that the constraints on the performance of *GrimK* strategies
 170 imposed by players' incentives and the presence of noise apply
 171 equally to any strategy in the trigger class.

		Noise (ε)				Level of Cooperation			
		0.1	0.05	0.01	0.001				
Survival Probability (γ)	0.85	0.8333	0.8846	0.8488	0.8412	0.85			
	0.9	0.8333	0.8354	0.8017	0.7944		0.8		
	0.95	0.8333	0.7915	0.7595	0.7526			0.75	
	0.99	0.8017	0.7595	0.7288	0.7222				0.7
	0.999	0.7944	0.7526	0.7222	0.7157				

Fig. 2. Upper bounds on cooperation. The entries are upper bounds on the share of cooperators possible in a *GrimK* equilibrium for various γ and ε values when $g = 0.5$ and $l = 2.5$, with a darker shade indicating a higher value as shown in the scale at right. As we move to the bottom right, the upper bound converges to $l/(1+l) \approx .7143$, which is the maximum share of cooperators sustainable in the limit, but away from the limit the upper bound can be different (the values in this table are all higher, but this is not the case for small γ or large ε).

172 **Stability, Convergence, and Evolutionary Properties.** *GrimK*
 173 strategies also satisfy desirable stability and convergence prop-
 174 erties. These derive from an important monotonicity property
 175 of *GrimK* strategies: when the distribution of individual
 176 records is more favorable today, the same will be true tomor-
 177 row, because players with better records both behave more
 178 cooperatively and induce more cooperative behavior from their
 179 partners. (See **Methods** for a precise statement.) From this
 180 observation it can be shown that, whenever the initial distri-
 181 bution of records is more favorable than the best steady-state
 182 record distribution, the record distribution converges to the
 183 best steady state. Similarly, whenever the initial distribution
 184 is less favorable than the worst steady state, convergence to
 185 the worst steady state obtains. See **Fig. 4**. These additional
 186 robustness properties are not shared by more complicated, non-
 187 monotone strategies that can sometimes support cooperation
 188 for a wider range of parameters than *GrimK*.

189 We also analyze evolutionary properties of *GrimK* equi-
 190 libria. When $g < l/(1+l)$, there is a sequence of *GrimK*
 191 equilibria that are “steady-state robust to mutants” and at-
 192 tains the maximum limit cooperation share of $l/(1+l)$. By
 193 this we mean that, when a small fraction of players adopt some
 194 mutant *GrimK'* strategy where $K' \neq K$, there is a steady-
 195 state distribution of records where it remains strictly optimal
 196 to play according to *GrimK*. We also perform simulations of
 197 dynamic evolution when a population playing a *GrimK* equi-
 198 librium is infected by a mutant population playing *GrimK'*
 199 for some $K' \neq K$. (See **Supplementary Information** and
 200 **Supplementary Fig. 1**.)

201 **Multplayer Public Goods Games.** Although our main analysis
 202 takes the basic unit of social interaction to be the standard
 203 2-player PD, many social interactions involve multiple players:
 204 the management of the commons and other public resources
 205 is a leading example (12, 13). In the **Supplementary Infor-**
 206 **mation** we establish that, when strategic complementarity

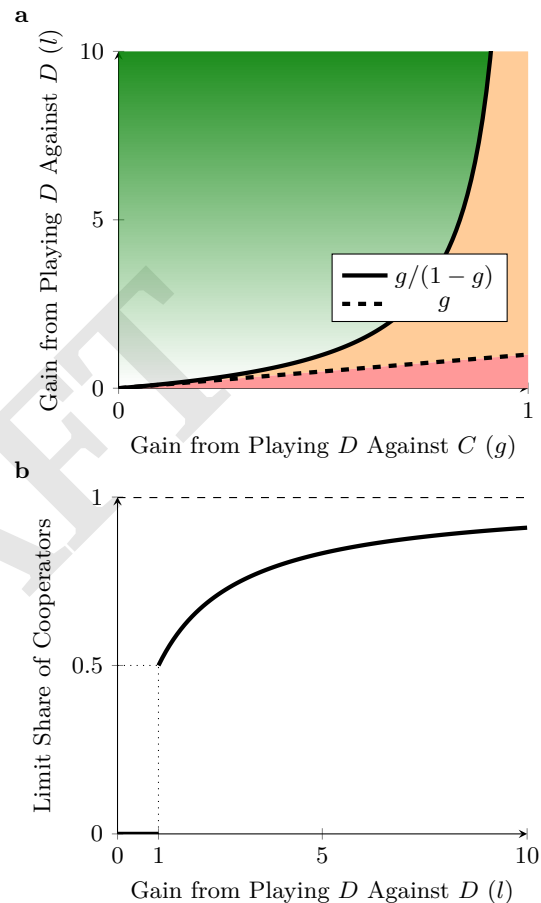


Fig. 3. Limit performance of *GrimK* strategies. **a**, In the green region ($l > g/(1-g)$), *GrimK* strategies sustain a positive limit share of cooperators, which increases with l , as indicated by a deeper shade of green. In the orange region ($g < l < g/(1-g)$), the limit share of cooperators with *GrimK* is 0, but other strategies may sustain positive cooperation in the limit. In the red region ($l \leq g$), individualistic records preclude cooperation. **b**, The limit share of cooperators as a function of l when $g = 1/2$. At $l = 1$, there is a discontinuity; as $l \rightarrow \infty$, the limit share of cooperators approaches 1.

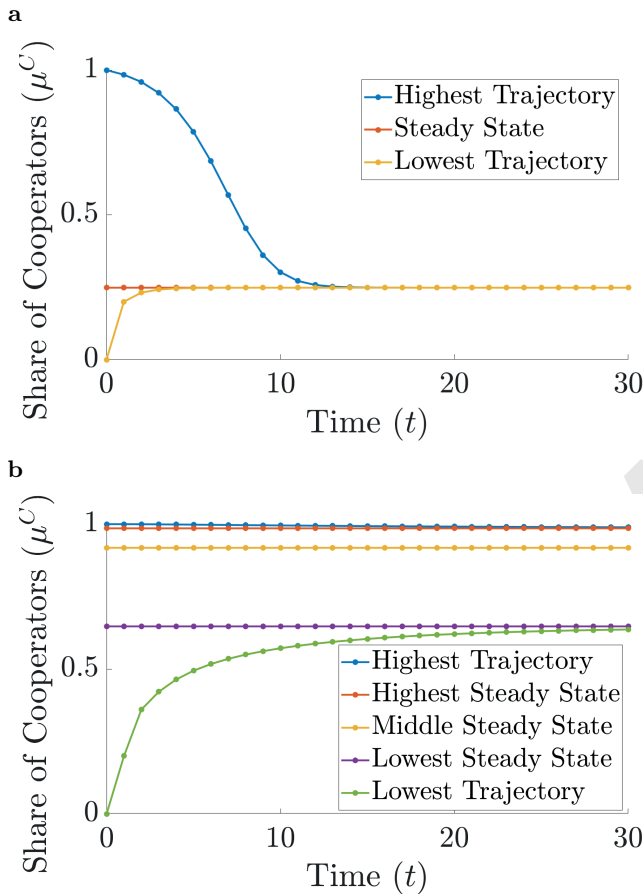


Fig. 4. Convergence of the share of cooperators. **a** depicts trajectories for the share of cooperators when $\gamma = 0.8$, $\varepsilon = 0.02$, and players use the *Grim1* strategy; **b** does the same for the *Grim2* strategy. **a**. All trajectories converge to the unique steady state; **b**, there are three steady states. Here “high” trajectories converge to the most cooperative steady state, while “low” trajectories converge to the least cooperative steady state. See **Methods** for details.

is sufficiently strong, robust cooperation in the multiplayer public goods game can be supported by a simple variant of *GrimK* strategies, wherein a player contributes to the public good if and only if all of their current partners have good records. In contrast, with strategic substitutability the unique strict equilibrium involves zero contribution. As the n -player public good game is a generalization of the PD, this implies that individualistic records preclude cooperation in the PD with strategic substitutability, as indicated in the red region in **Fig. 3a**.

Discussion

We have shown how individualistic records robustly support indirect reciprocity in supermodular PD and multiplayer public goods games. To place our results in context, recall that scoring models do not provide robust incentives, while standing models compute records as a recursive function of a player’s partners’ past actions and standing, their partners’ actions and standing, and so on, and thus require more information than may typically be available. The simplicity and power of individualistic records suggests that they may be usefully adapted to specific settings where cooperation is based on indirect reciprocity, such as online rating systems (24, 25), credit ratings (10, 26), decentralized currencies (27, 28), and monitoring systems for conflict resolution (29). Individualistic records may also prove useful in modeling the role of costly punishment in the evolution of cooperation (30–33).

We interpret individualistic records and *GrimK* strategies as both a theoretical demonstration that simple strategies can sometimes support cooperation using only first-order information and as an approximation of human behavior in a range of environments. For example, when meeting a potential business partner for the first time, it is common to contact their past partners and inquire about the potential partner’s past behavior, typically without delving into the past partners’ own past behavior or the past partners’ partners’ behavior. Similarly, in online marketplaces such as Ebay or AirBnB, one typically rates one’s current partner’s behavior in the absence of any information about their own past partners’ behavior. Users then observe summary statistics that depend only on their current partner’s own past behavior, which is an example of individualistic (first-order) records. Moreover, if users behave honestly only with partners who have not received too many negative reviews, their behavior can be approximated by *GrimK* strategies.

We conclude by discussing possible extensions of our analysis.

First, while we have analyzed the evolutionary stability of the *GrimK* equilibrium, we have not analyzed how this equilibrium could first arise. In our model, it is a strict equilibrium for all agents to always defect, so that equilibrium is also an ESS. To explain how society might move from such a state to a more cooperative equilibrium such as *GrimK*, we could appeal to random mutations. Given our continuum population, this could be modeled as a deterministic drift as in (34), but we do not develop that argument here.

We have also assumed that everyone shares the same assessment of each individual’s record. This “public information” assumption is known to be critical in some prior models of indirect reciprocity. In our model, allowing heterogeneous assessments of a player’s record would not change the analysis

267 very much, so long as both partners learn their opponents' 268
 268 assessments of their records before taking actions (35–37). The 269
 269 more complex situation where each partner's assessment of 270
 270 the other's record is private information would be interesting 271
 271 to study in future research.

272 Methods

273 Here we summarize the model and mathematical results; fur- 274
 274 ther details are provided in the **Supplementary Informa-**
 275 **tion.**

276 **A Model of Social Cooperation with Individualistic**
 277 **Records.** Time is discrete and doubly infinite:
 278 $t \in \{\dots, -2, -1, 0, 1, 2, \dots\}$. There is a population of
 279 individuals of unit mass, each with survival probability
 280 $\gamma \in (0, 1)$, so each individual's lifespan is geometrically
 281 distributed with mean $1/(1 - \gamma)$. An inflow of $1 - \gamma$ newborn
 282 players each period keeps the total population size constant.
 283 We thus have an infinite-horizon dynamic model with
 284 overlapping generations of players (38).

285 Every period, individuals randomly match in pairs to play
 286 the PD (**Fig. 1**). Each individual carries a *record* $k \in \mathbb{N} :=$
 287 $\{0, 1, 2, \dots\}$. Newborns have record 0. Under the Counting
 288 D 's record system, whenever an individual plays D , their
 289 record increases by 1, while whenever an individual plays
 290 C , their record remains constant with probability $1 - \varepsilon$ and
 291 increases by 1 with probability ε ; thus, $\varepsilon \in (0, 1)$ measures
 292 the amount of noise in the system (39–43). More generally, a
 293 record system specifies an arbitrary next-period record as a
 294 function of the current-period record and the current-period
 295 *recorded action*, which equals D if the individual plays D , and
 296 equals C with probability $1 - \varepsilon$ and equals D with probability
 297 ε if the individual plays C .

298 When two players meet, they observe each other's records
 299 and nothing else. A *strategy* is a mapping $\mathbf{s} : \mathbb{N} \times \mathbb{N} \rightarrow \{C, D\}$,
 300 with the convention that the first component of the domain is
 301 a player's own record and the second component is the current
 302 opponent's record. We assume that all players use the same
 303 strategy, noting that this must be the case in every strict
 304 equilibrium in a symmetric, continuum-agent model like ours.
 305 (Of course, players who have different records and/or meet
 306 opponents with different records may take different actions.)

307 The *state* of the system $\mu \in \Delta(\mathbb{N})$ describes the share of
 308 the population with each record, where $\mu_k \in [0, 1]$ denotes
 309 the share with record k . When all players use strategy \mathbf{s} , let
 310 $f_{\mathbf{s}} : \Delta(\mathbb{N}) \rightarrow \Delta(\mathbb{N})$ denote the resulting *update map* govern-
 311 ing the evolution of the state. (The formula for $f_{\mathbf{s}}(\mu)$ is in
 312 the **Supplementary Information**.) A *steady state* under
 313 strategy \mathbf{s} is a state μ such that $f_{\mathbf{s}}(\mu) = \mu$.

314 Given a strategy \mathbf{s} and state μ , the expected flow payoff of a
 315 player with record k is $\pi_k(\mathbf{s}, \mu) = \sum_{k'} \mu_{k'} u(\mathbf{s}(k, k'), \mathbf{s}(k', k))$,
 316 where u is the PD payoff function. Denote the probabil-
 317 ity that a player with current record k has record k' t
 318 periods in the future by $\phi_k(\mathbf{s}, \mu)^t(k')$. The continuation
 319 payoff of a player with record k is then $V_k(\mathbf{s}, \mu) = (1 -$
 320 $\gamma) \sum_{t=0}^{\infty} \gamma^t \sum_{k'} \phi_k(\mathbf{s}, \mu)^t(k') \pi_{k'}(\mathbf{s}, \mu)$. Note that we have nor-
 321 malized continuation payoffs by $(1 - \gamma)$ to express them in
 322 per-period terms. A player's objective is to maximize their
 323 expected lifetime payoff.

324 A pair (\mathbf{s}, μ) is an *equilibrium* if μ is a steady-state un-
 325 der \mathbf{s} and, for each own record k and opponent's record k' ,

the prescribed action $\mathbf{s}(k, k') \in \{C, D\}$ maximizes the ex- 326
 326 pected lifetime payoff from the current period onward, given 327
 327 by $(1 - \gamma)u(a, \mathbf{s}(k', k)) + \gamma \sum_{k''} (\rho(k, a)[k'']) V_{k''}(\mathbf{s}, \mu)$, over 328
 328 $a \in \{C, D\}$, where $\rho(k, a)[k'']$ denotes the probability that a 329
 329 player with record k who takes action a acquires next-period 330
 330 record k'' . Note that this expression depends on the opponent's 331
 331 record only through the predicted current-period opponent 332
 332 action, $\mathbf{s}(k', k)$. In addition, the ratio $(1 - \gamma)/\gamma$ captures the 333
 333 weight that players place on their current payoff relative to 334
 334 their continuation payoff from tomorrow on. We study limits 335
 335 where this ratio converges to 0, as opposed to time-average 336
 336 payoffs which give exactly 0 weight to any one period's payoff, 337
 337 because in the latter case optimization and equilibrium impose 338
 338 unduly weak restrictions (44). An equilibrium is *strict* if the 339
 339 maximizer is unique for all pairs (k, k') , i.e. the optimal action 340
 340 is always unique. Note that this equilibrium definition allows 341
 341 agents to maximize over all possible strategies, as opposed to 342
 342 only strategies from some pre-selected set. We focus on 343
 343 strict equilibria because they are robust: they remain equi- 344
 344 libria under “small” perturbations of the model. Note that 345
 345 the strategy *Always Defect*, i.e. $\mathbf{s}(k, k') = D$ for all (k, k') , 346
 346 together with any steady state is always a strict equilibrium. 347
 347 Lemma 2 in the **Supplementary Information** character- 348
 348 izes the steady states for any *GrimK* strategy, as well as the 349
 349 $\gamma, \varepsilon, g, l$ parameters for which the steady states are equilibria. 350

351 **Limit Cooperation under GrimK Strategies.** Under *GrimK* 351
 352 strategies, a matched pair of players cooperate if and only 352
 352 if both records are below a pre-specified cutoff K : that 353
 353 is, $\mathbf{s}(k, k') = C$ if $\max\{k, k'\} < K$, and $\mathbf{s}(k, k') = D$ if 354
 354 $\max\{k, k'\} \geq K$. 355

356 We call an individual a *cooperator* if their record is below 356
 356 K and a *defector* otherwise. Note that each individual may be 357
 357 a cooperator for some periods of their life and a defector for 358
 358 other periods, rather than being pre-programmed to cooperate 359
 359 or defect for their entire life. 360

361 Given an equilibrium strategy *GrimK*, let $\mu^C = \sum_{k=0}^{K-1} \mu_k$ 361
 362 denote the corresponding steady-state share of cooperators. 362
 362 Note that, in a steady state with cooperator share μ^C , mutual 363
 363 cooperation is played in share $(\mu^C)^2$ of all matches. Let 364
 364 $\bar{\mu}^C(\gamma, \varepsilon)$ be the maximal share of cooperators in any tolerant 365
 365 grim trigger equilibrium (allowing for every possible K) when 366
 366 the survival probability is γ and the noise level is ε . 367

368 Theorem 1 in the **Supplementary Information** charac- 368
 368 terizes the performance of equilibria in *GrimK* strategies in 369
 369 the double limit where the survival probability approaches 1— 370
 370 so that players expect to live a long time and the “shadow of 371
 371 the future” looms large—and the noise level approaches 0—so 372
 372 that records are reliable enough to form the basis for incentives. 373
 373 (This long-lifespan/low-noise limit is the leading case of inter- 374
 374 est in theoretical analyses of indirect reciprocity (8, 45–49).) 375
 375 The theorem shows that, in the double limit $(\gamma, \varepsilon) \rightarrow (1, 0)$, 376
 376 $\bar{\mu}^C(\gamma, \varepsilon)$ converges to $l/(1+l)$ when $g < l/(1+l)$, and converges 377
 377 to 0 when $g > l/(1+l)$. The formal statement and proof of this 378
 378 result are contained in the **Supplementary Information**. 379

380 Barring knife-edge cases, tolerant grim trigger strategies can 380
 381 thus robustly support positive cooperation in the double limit 381
 382 $(\gamma, \varepsilon) \rightarrow (1, 0)$ if and only if the gain from defecting against a 382
 382 partner who cooperates is significantly smaller than the loss 383
 383 from cooperating against a partner who defects: $g < l/(1+l)$. 384
 384 Moreover, the maximum level of cooperation in this case is 385
 385 $l/(1+l)$. Here we explain the logic of this result. 386

We first show that $g < \mu^C$ in any *GrimK* equilibrium. Newborn individuals have continuation payoff equal to the average payoff in the population, which is $(\mu^C)^2$. Thus, since a newborn player plays C if and only if matched with a cooperator, $(\mu^C)^2 = (1 - \gamma)\mu^C + \gamma\mu^C V_0^C + \gamma(1 - \mu^C)V_0^D$, where V_0^C and V_0^D are the expected continuation payoffs of a newborn player after playing C and D , respectively. Newborn players have the highest continuation payoff in the population, so $V_0^C \leq V_0 = (\mu^C)^2$. For a newborn player to prefer not to cheat a cooperative partner, it must be that $V_0^D < V_0^C - (1 - \gamma)g/\gamma$, so when $\mu^C < 1$ (as is necessarily the case with any noise),

$$(\mu^C)^2 < (1 - \gamma)\mu^C + \gamma(\mu^C)^2 - (1 - \gamma)(1 - \mu^C)g.$$

This inequality can hold only if $g < \mu^C$.

We next show that $\gamma(1 - \varepsilon)\mu^C < l/(1 + l)$ in any *GrimK* equilibrium. The continuation payoff V_{K-1} of an individual with record $K - 1$ satisfies $V_{K-1} = (1 - \gamma)\mu^C + \gamma(1 - \varepsilon)\mu^C V_{K-1}$, or $V_{K-1} = (1 - \gamma)\mu^C / (1 - \gamma(1 - \varepsilon)\mu^C)$. A necessary condition for an individual with record $K - 1$ to prefer to play D against a defector partner is $(1 - \gamma)(-l) + \gamma(1 - \varepsilon)V_{K-1} < 0$, or $l > \gamma(1 - \varepsilon)V_{K-1}/(1 - \gamma)$. Combining this inequality with the expression for V_{K-1} yields $\gamma(1 - \varepsilon)\mu^C < l/(1 + l)$, which in the $(\gamma, \varepsilon) \rightarrow (1, 0)$ limit gives $\mu^C \leq l/(1 + l)$.

We have established that tolerant grim trigger strategies can support positive cooperation in the $(\gamma, \varepsilon) \rightarrow (1, 0)$ limit only if $g \leq l/(1 + l)$, and that the maximum cooperation share cannot exceed $l/(1 + l)$. The proof of Theorem 1 is completed by showing that when $g < l/(1 + l)$, by carefully choosing the tolerance level K , *GrimK* can support cooperation shares arbitrarily close to any value between g and $l/(1 + l)$ in equilibrium when the survival probability is close to 1 and the noise level is close to 0.

Limit Cooperation under General Trigger Strategies. *GrimK* strategies are an instance of the more general class of *trigger strategies*, which are defined by the following properties: (i) The set of all possible records can be partitioned into two classes, “good records” G and “bad records” B . (ii) Partners cooperate if and only if they both have good records: $s(k, k') = C$ for all pairs $(k, k') \in G \times G$, and $s(k, k') = D$ for all other pairs (k, k') . (iii) The class B is absorbing: if $k \in B$, then every record k' that can be reached starting at record k is also in B .

Theorem 9 in the **Supplementary Information** shows that, in the $(\gamma, \varepsilon) \rightarrow (1, 0)$ double limit, the maximum steady-state share of good-record players that can be supported in any trigger strategy equilibrium converges to zero if $g > l/(1 + l)$, and converges to $l/(1 + l)$ if $g < l/(1 + l)$. Thus, in this double limit, tolerant grim trigger strategies attain the most equilibrium cooperation that any trigger strategy can support.

The intuition for this result is that the necessary conditions $g < \mu^C$ and $\gamma(1 - \varepsilon)\mu^C < l/(1 + l)$ derived above for *GrimK* strategies apply equally to any trigger strategy. The argument to establish the necessity of $g < \mu^C$ is similar to that for *GrimK* strategies, except we must now consider the incentives of a player with whichever record k yields the greatest equilibrium continuation payoff, which is no longer necessarily a newborn (i.e., we may now have $k \neq 0$). The argument to establish necessity of $\gamma(1 - \varepsilon)\mu^C < l/(1 + l)$ is also similar to that for *GrimK* strategies, but now we consider the incentives

of any player with a “marginal” good record that will become bad if the player is recorded as playing one additional D , which is no longer necessarily a player who has been recorded as playing $K - 1$ D 's for some fixed cutoff K .

Convergence of *GrimK* Strategies. Fix an arbitrary initial record distribution $\mu^0 \in \Delta(\mathbb{N})$. When all individuals use *GrimK* strategies, the population share with record k at time t , μ_k^t , evolves according to

$$\begin{aligned} \mu_0^{t+1} &= 1 - \gamma + \gamma(1 - \varepsilon)\mu^{C,t}\mu_0^t, \\ \mu_k^{t+1} &= \gamma(1 - (1 - \varepsilon)\mu^{C,t})\mu_{k-1}^t + \gamma(1 - \varepsilon)\mu^{C,t}\mu_k^t \text{ for } 0 < k < K, \end{aligned}$$

where $\mu^{C,t} = \sum_{k=0}^{K-1} \mu_k^t$.

Fixing K , we say that distribution μ *dominates* (or is *more favorable than*) distribution $\tilde{\mu}$ if, for every $k < K$, $\sum_{\bar{k}=0}^k \mu_{\bar{k}} \geq \sum_{\bar{k}=0}^k \tilde{\mu}_{\bar{k}}$; that is, if for every $k < K$ the share of the population with record no worse than k is greater under distribution μ than under distribution $\tilde{\mu}$. Under the *GrimK* strategy, let $\bar{\mu}$ denote the steady state with the largest share of cooperators, and let $\underline{\mu}$ denote the steady state with the smallest share of cooperators.

Theorem 12 in the **Supplementary Information** shows that, if the initial record distribution is more favorable than $\bar{\mu}$, then the record distribution converges to $\bar{\mu}$; similarly, if the initial record distribution is less favorable than $\underline{\mu}$, then the record distribution converges to $\underline{\mu}$. Formally, if μ^0 dominates $\bar{\mu}$, then $\lim_{t \rightarrow \infty} \mu^t = \bar{\mu}$; similarly, if μ^0 is dominated by $\underline{\mu}$, then $\lim_{t \rightarrow \infty} \mu^t = \underline{\mu}$.

In **Fig. 4a** the blue trajectory corresponds to the initial distribution where all players have record 0, the red trajectory is constant at the unique steady-state value $\mu^C \approx .2484$, and the yellow trajectory corresponds to the initial distribution where all players have defector records. Here all the trajectories converge to the unique steady state. In **Fig. 4b**, the red trajectory is constant at the largest steady-state value $\mu^C \approx .9855$, the yellow trajectory is constant at the intermediate steady-state value $\mu^C \approx .9184$, and the purple trajectory is constant at the smallest steady-state value $\mu^C \approx .6471$. The blue trajectory corresponds to the initial distribution where all players have record 0 and converges to the largest steady-state share of cooperators. The green trajectory corresponds to the initial distribution where all players have defector records and converges to the smallest steady-state share of cooperators.

Code availability. All simulations and numerical calculations have been performed with MATLAB R2017b and Wolfram Mathematica 11.3.0.0. In the Appendix of the **Supplementary Information**, we provide the MATLAB scripts used to generate **Fig. 4** as well as those to simulate evolutionary dynamics and generate **Supplementary Fig. 1**.

ACKNOWLEDGMENTS. This work was supported by National Science Foundation grants SES-1643517 and SES-1555071 and Sloan Foundation grant 2017-9633.

1. MA Nowak, K Sigmund, Evolution of indirect reciprocity by image scoring. *Nature* **393**, 573–577 (1998).
2. MA Nowak, K Sigmund, The dynamics of indirect reciprocity. *Journal of Theoretical Biology* **194**, 561–574 (1998).
3. R Sugden, New Developments in the Theory of Choice Under Uncertainty. *Bulletin of Economic Research* **38**, 1–24 (1986).
4. M Kandori, Social Norms and Community Enforcement. *The Review of Economic Studies* **59**, 63 (1992).

- 485 5. O Leimar, P Hammerstein, Evolution of cooperation through indirect reciprocity. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **268**, 745–753 (2001).
- 486 6. K Panchanathan, R Boyd, A tale of two defectors: the importance of standing for evolution of indirect reciprocity. *Journal of Theoretical Biology* **224**, 115–126 (2003).
- 487 7. MA Nowak, K Sigmund, Evolution of indirect reciprocity. *Nature* **437**, 1291 (2005).
- 488 8. S Takahashi, Community enforcement when players observe partners' past play. *Journal of Economic Theory* **145**, 42–62 (2010).
- 489 9. Y Heller, E Mohlin, Observations on Cooperation. *Review of Economic Studies*, 2253–2282 (2018).
- 490 10. V Bhaskar, C Thomas, Community Enforcement of Trust with Bounded Memory. *Review of Economic Studies* (2018).
- 491 11. B Klein, KB Leffler, The role of market forces in assuring contractual performance. *Journal of Political Economy* **89**, 615–641 (1981).
- 492 12. G Hardin, The tragedy of the commons. *Science* **162**, 1243–1248 (1968).
- 493 13. E Ostrom, *Governing the commons: The evolution of institutions for collective action*. (Cambridge University Press), (1990).
- 494 14. S Coate, M Ravallion, Reciprocity without commitment: Characterization and performance of informal insurance arrangements. *Journal of Development Economics* **40**, 1–24 (1993).
- 495 15. LA Dugatkin, *Cooperation among animals: an evolutionary perspective*. (Oxford University Press on Demand), (1997).
- 496 16. K Sigmund, *The calculus of selfishness*. (Princeton University Press) Vol. 6, (2010).
- 497 17. R Axelrod, WD Hamilton, The evolution of cooperation. *Science (New York, N.Y.)* **211**, 1390–6 (1981).
- 498 18. JI Bulow, JD Geanakoplos, PD Klemperer, Multimarket oligopoly: Strategic substitutes and complements. *Journal of Political Economy* **93**, 488–511 (1985).
- 499 19. D Fudenberg, J Tirole, The fat-cat effect, the puppy-dog ploy, and the lean and hungry look. *American Economic Review* **74**, 361–366 (1984).
- 500 20. A Zahavi, Altruism as a handicap: the limitations of kin selection and reciprocity. *Journal of Avian Biology* **26**, 1–3 (1995).
- 501 21. JW Friedman, A non-cooperative equilibrium for supergames. *The Review of Economic Studies* **38**, 1–12 (1971).
- 502 22. JM Smith, *Evolution and the Theory of Games*. (Cambridge University Press), (1982).
- 503 23. JW Weibull, *Evolutionary Game Theory*. (MIT Press), (1997).
- 504 24. P Resnick, K Kuwabara, R Zeckhauser, E Friedman, Reputation systems. *Communications of the ACM* **43**, 45–48 (2000).
- 505 25. C Dellarocas, Reputation mechanism design in online trading environments with pure moral hazard. *Information Systems Research* **16**, 209–230 (2005).
- 506 26. DB Klein, Promise keeping in the great society: A model of credit information sharing. *Economics & Politics* **4**, 117–136 (1992).
- 507 27. N Kocherlakota, N Wallace, Incomplete record-keeping and optimal payment arrangements. *Journal of Economic Theory* **81**, 272–289 (1998).
- 508 28. B Biais, C Bisiere, M Bouvard, C Casamatta, The blockchain folk theorem. *The Review of Financial Studies* **32**, 1662–1715 (2019).
- 509 29. JD Fearon, DD Laitin, Explaining interethnic cooperation. *American Political Science Review* **90**, 715–735 (1996).
- 510 30. E Fehr, S Gächter, Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives* **14**, 159–181 (2000).
- 511 31. R Bhui, M Chudek, J Henrich, How exploitation launched human cooperation. *Behavioral Ecology and Sociobiology* **73**, 78 (2019).
- 512 32. R Boyd, H Gintis, S Bowles, PJ Richerson, The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences* **100**, 3531–3535 (2003).
- 513 33. J Henrich, et al., Costly punishment across human societies. *Science* **312**, 1767–1770 (2006).
- 514 34. D Fudenberg, C Harris, Evolutionary dynamics with aggregate shocks. *Journal of Economic Theory* **57**, 420–441 (1992).
- 515 35. S Uchida, Effect of private information on indirect reciprocity. *Physical Review E* **82**, 036111 (2010).
- 516 36. C Hilbe, L Schmid, J Tkadlec, K Chatterjee, MA Nowak, Indirect reciprocity with private, noisy, and incomplete information. *Proceedings of the National Academy of Sciences* **115**, 12241–12246 (2018).
- 517 37. H Ohtsuki, Y Iwasa, MA Nowak, Reputation effects in public and private interactions. *PLoS computational biology* **11** (2015).
- 518 38. D Fudenberg, K He, Learning and type compatibility in signaling games. *Econometrica* **86**, 1215–1255 (2018).
- 519 39. S Le, R Boyd, Evolutionary dynamics of the continuous iterated prisoner's dilemma. *Journal of Theoretical Biology* **245**, 258–267 (2007).
- 520 40. D Fudenberg, E Maskin, Evolution and cooperation in noisy repeated games. *American Economic Review: Papers and Proceedings* **80**, 274 (1990).
- 521 41. D Fudenberg, DG Rand, A Dreber, Slow to Anger and Fast to Forgive: Cooperation in an Uncertain World. *American Economic Review* **102**, 720–749 (2012).
- 522 42. JM McNamara, Z Barta, AI Houston, Variation in behaviour promotes cooperation in the prisoner's dilemma game. *Nature* **428**, 745 (2004).
- 523 43. J Bendor, RM Kramer, S Stout, When in doubt... cooperation in a noisy prisoner's dilemma. *Journal of Conflict Resolution* **35**, 691–719 (1991).
- 524 44. D Fudenberg, E Maskin, The folk theorem in repeated games with discounting or with incomplete information. *Econometrica* **54**, 533 (1986).
- 525 45. H Ohtsuki, Y Iwasa, How should we define goodness?—reputation dynamics in indirect reciprocity. *Journal of Theoretical Biology* **231**, 107–120 (2004).
- 526 46. H Ohtsuki, Y Iwasa, The leading eight: social norms that can maintain cooperation by indirect reciprocity. *Journal of Theoretical Biology* **239**, 435–444 (2006).
- 527 47. H Brandt, K Sigmund, Indirect reciprocity, image scoring, and moral hazard. *Proceedings of the National Academy of Sciences* **102**, 2666–2670 (2005).
- 528 48. G Ellison, Cooperation in the prisoner's dilemma with anonymous random matching. *The Review of Economic Studies* **61**, 567–588 (1994).
- 529 49. J Hörner, W Olszewski, The folk theorem for games with private almost-perfect monitoring. *Econometrica* **74**, 1499–1544 (2006).

Figure Legends

Legend for Figure 1. The prisoner's dilemma. The matrices on the left show how any prisoner's dilemma can be represented by the standard normalization with $g = (T - R)/(R - P)$ and $l = (P - S)/(R - P)$, where $T > R > P > S$. The matrices on the right illustrate this normalization for “donation games” in which choosing G (*Give*) instead of S (*Shirk*) incurs a personal cost c and gives benefit $b > c$ to the opponent.

Legend for Figure 2. Upper bounds on cooperation. The entries are upper bounds on the share of cooperators possible in a *GrimK* equilibrium for various γ and ε values when $g = 0.5$ and $l = 2.5$, with a darker shade indicating a higher value as shown in the scale at right. As we move to the bottom right, the upper bound converges to $l/(1+l) \approx .7143$, which is the maximum share of cooperators sustainable in the limit, but away from the limit the upper bound can be different (the values in this table are all higher, but this is not the case for small γ or large ε).

Legend for Figure 3. Limit performance of *GrimK* strategies. **a**, In the green region ($l > g/(1-g)$), *GrimK* strategies sustain a positive limit share of cooperators, which increases with l , as indicated by a deeper shade of green. In the orange region ($g < l < g/(1-g)$), the limit share of cooperators with *GrimK* is 0, but other strategies may sustain positive cooperation in the limit. In the red region ($l \leq g$), individualistic records preclude cooperation. **b**, The limit share of cooperators as a function of l when $g = 1/2$. At $l = 1$, there is a discontinuity; as $l \rightarrow \infty$, the limit share of cooperators approaches 1.

Legend for Figure 4. Convergence of the share of cooperators. **a** depicts trajectories for the share of cooperators when $\gamma = 0.8$, $\varepsilon = 0.02$, and players use the *Grim1* strategy; **b** does the same for the *Grim2* strategy. **a**, All trajectories converge to the unique steady state; **b**, there are three steady states. Here “high” trajectories converge to the most cooperative steady state, while “low” trajectories converge to the least cooperative steady state. See **Methods** for details.