Journal of Economic Behavior and Organization 167 (2019) 228–234

Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

journal homepage: www.elsevier.com/locate/jebo

Modelling kindness☆

Martin Dufwenberg[a], Georg Kirchsteiger[b,*]

[a] University of Arizona, University of Gothenburg, CESifo, Sweden
[b] Universite Libre de Bruxelles, Belgium

ARTICLE INFO

Article history:
Received 27 March 2018
Revised 25 June 2018
Accepted 27 July 2018
Available online 1 December 2018

JEL classification:
C70
D01
D91

Keywords:
Kindness
Reciprocity
Efficient strategies

ABSTRACT

Kindness is an important concept in reciprocity theory, and may matter also for other forms of motivation. We critically compare definitions proposed by Rabin (1993) and by Dufwenberg and Kirchsteiger (2004). Several reasons to prefer the latter definition are highlighted, but also a flaw (discovered by Isoni and Sugden 2018) which we show can be eliminated using a slightly revised definition.

1. Introduction

Kindness may shape goals and decisions: People may enjoy being kind. Or they may like it when others view them as kind. Or they may wish to be kind in return.

While kindness may thus plausibly matter for many reasons, the third one we listed – concerning kindness-based reciprocity – is what economists have focused on. In a seminal paper, (Rabin, 1993) developed the first explicit kindness notion in order to model reciprocal behavior. See Dufwenberg and Kirchsteiger (2004) (D&K), Falk and Fischbacher (2006), Segal and Sobel (2007), Sobel (2005), and Sebald (2010) for examples of follow-up work.[1]

We offer critical reflections on how to best model kindness. We focus on kindness in reciprocity theory, but believe that the right definitions for that context will likely be viable for other purposes too.

The key issues surround how Rabin's and D&K's kindness definitions differ in some subtle but important ways. Section 2 sets the stage, presenting backdrop issues and definitions. Section 3 points to three properties that we believe

☆ We have benefited from spirited exchanges with Senran Lin, Amrish Patel, Bob Sugden, Claudia Toma, and several referees (also on other papers).

* Corresponding author.
E-mail addresses: martind@eller.arizona.edu (M. Dufwenberg), gkirchst@ulb.ac.be (G. Kirchsteiger).

[1] More applied work involving reciprocity theory (especially D&K's model) studied wage setting (Dufwenberg and Kirchsteiger, 2000), voting (Hahn, 2009), framing effects (Dufwenberg et al., 2011), hold-up (Dufwenberg et al., 2013), ultimatum bargaining (van Damme et al., 2014, section 6, by D&K), gift exchange (Netzer and Schmutzler, 2014), mechanism design (Bierbrauer and Netzer, 2016; Bierbrauer et al., 2017; Netzer and Volk, 2014), co-financing agreements (Jang et al.2016), insolvency in banking (Dufwenberg and Rietzke, 2016), trade disputes (Conconi et al., 2017), public goods (Dufwenberg and Patel, 2017), randomized controlled trials (Aldashev et al., 2017), climate negotiations (Nyborg, 2018), and communication (Le Quement and Patel, 2018).

https://doi.org/10.1016/j.jebo.2018.07.014
0167-2681/

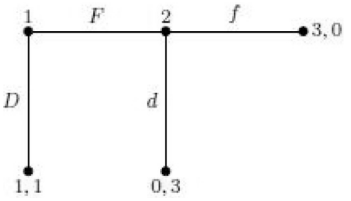
speak in favor of D&K's formulation. Isoni and Sugden (2018) (I&S) point out a flaw though, and in Section 4 we present a revised definition that overcomes the issue. Section 5 concludes.

2. Preliminaries

Take a game form with the usual definition of strategies. A strategy combination resulting from the individual strategy choices leads to a material payoff like money for each player. Within this framework, Rabin defines the kindness of an actual strategy choice of player i vis-a-vis player j as the material payoff i's choice gives j, given i's belief about the other players' strategies, minus the so-called "equitable payoff," i.e. the average between the minimum and the maximum i can give to j, given i's beliefs.[2]

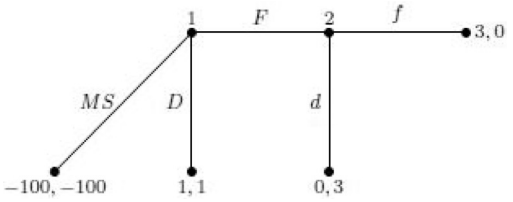
This notion is based on two crucial insights. First, the kindness of a strategy choice of player i vis-a-vis player j depends not only on the absolute material payoff i's strategy choice gives to j, but also on the range of material payoffs i could have given to j. To see this, take an individual decision problem where "player 1" has to choose between D and F. D leads to an allocation of material payoffs of (1,1) for 1 and another agent, player 2. Is player 1's the choice of D kind? It obviously depends on the material payoffs that would result from choice F. If F leads to an allocation (3,0), 1's choice of D is kind, since it gives player 2 a payoff of 1 instead of 0. But if the alternative allocation would be (0,3), D is an unkind action. Hence, the kindness of a player i's choice depends on the material payoffs he could have given to player j.

Second, in a strategic situation the kindness of a player depends on his beliefs about the other players' strategy choices, and the kindness of a player as perceived by the others depends on the second order beliefs of the others. To see this look the following Example 1 (akin to one of D&K's):



Is F an unkind action? Clearly, this depends on what player 1 believes that player 2 will do. Suppose 1 believes that 2 will choose d. By choosing F player 1 then intends to give a payoff of 3 to player 2, whereas player 2 would get a payoff of only 1 if player 1 chose D. Hence, one may conclude that player 1 acts kindly if he chooses F. By an analogous argument, however, one must conclude that 1 is unkind if he chooses F while believing that 2 will choose f. Furthermore, to evaluate the kindness of another player, second order beliefs are crucial: In order to evaluate 1's kindness, player 2 must form a belief about 1's beliefs about 2's choice.

These two issues – that i's kindness has to judged by the actual material payoff i gives to j, relative to the payoff feasible to j, and that kindness and perceived kindness depends on beliefs – are taken care of by the notions described above. But there is a problem with these. Take Example 2, which is the same as before except that now player 1 has an additional strategy MS that gives both players a material payoff of -100.



Adding such a "murder/suicide" option to the decision problem would make all other choices of player 1 kind if one takes kindness as defined above, for any feasible beliefs about 2's choice. And 2 would regard 1's choices of D and of F as kind for any second order beliefs 2 might hold. But the feasibility of this horrific choice MS seems to be irrelevant for drawing conclusions regarding the kindness of the choices F and D.

In the literature one finds two ways how to deal with this issue. On the one hand, Rabin amends his kindness definition such that for the calculation of the equitable payoff only those strategies of player i are taken into account that do not lead to outcomes that are Pareto-dominated (in terms of material payoffs) by another outcome that would be feasible given i's beliefs about the other players' strategies. With this approach the addition of the MS option in example 2 does not change the equitable payoff, and hence does not change the kindness of D and F.

D&K propose a different way to deal with the problem. For the calculation of the equitable payoff they take into account only those strategies of i that do not for sure lead to Pareto-inferior outcomes in material terms. Obviously, this also excludes MS, and hence its inclusion does not change the evaluation of the kindness of strategies D and F.

[2] Rabin uses a normalization of the kindness function such that the most kind strategy has always a kindness of 1/2, and the most unkind one of -1 (unless the two coincide). For the purpose of this paper, we ignore this normalization since our arguments would not change if we took it into account.

While these notions apply similarly to Example 2, this is not true in general. To highlight the issues,[3] we have to introduce more notation. Take a finite two-player extensive game form with perfect information. We mostly restrict attention to such games because while extensions are straightforward they would require additional notation without providing additional insights (see D&K for the general case, as well as footnote 10 below).

A_i denotes the set of behavioral strategies of player i, with a_i being a typical element of A_i. Denote by $\pi_i : \prod_{j\in I} A_j \to \mathbb{R}$ player i's (expected) material payoff function. κ_{ij} denotes how kind player i is to j. λ_{iji} denotes player i's belief about how kind j is to i. Denote by $b_{ij} \in A_j$ the first-order belief of player i about player j's strategy. $c_{iji} \in A_i$ is the second-order belief of i about j's belief about the i's strategy.[4] The payoff player i intends to give j is given by $\pi_j(a_i, b_{ij})$. The equitable payoff, π_j^e, i.e. the average payoff i thinks that he could give to j, is given by

$$\pi_j^e(b_{ij}) = \frac{1}{2}\left(\max_{a_i\in A_i} \pi_j(a_i, b_{ij}) + \min_{a_i\in E_i} \pi_j(a_i, b_{ij})\right),$$

with $E_i \subseteq A_i$ being what Rabin and D&K refer to as the set of "efficient strategies" although their definitions differ as we'll explain shortly. Player i's kindness from choosing a_i when holding belief b_{ij}, is defined as

$$\kappa_{ij}\left(a_i, b_{ij}\right) = \pi_j(a_i, b_{ij}) - \pi_j^e(b_{ij}).$$

The linear structure of the equitable payoff function and of the kindness function are taken from D&K and used for ease of exposition. But it can be easily checked that all our arguments and examples are valid for any other functional form as long as it implies that kindness strictly decreases in $\min_{a_i\in E_i} \pi_j(a_i, b_{ij})$.[5]

Player j's belief about the kindness of player i, λ_{jij}, is derived the same way as κ_{ij}, replacing i's strategy a_i by b_{ji} (j's first-order belief about i's strategy), and replacing i's first-order belief b_{ij} about j's strategy by c_{jij} (j's second-order belief about i's first-order belief about j's strategy). Formally

$$\lambda_{jij}\left(b_{ji}, c_{jij}\right) = \pi_j(b_{ji}, c_{jij}) - \pi_j^e(c_{jij}).$$

Rabin defines $E_i = E_i^{Rabin}(b_{ij})$, a set that depends on i's beliefs via b_{ij}:

$$E_i^{Rabin}(b_{ij}) = \left\{ a_i \in A_i \,\middle|\, \begin{array}{c} \nexists a_i' \in A_i \text{ with } \pi_k(a_i', b_{ij}) \geq \pi_k(a_i, b_{ij}) \text{ for all } k \in \{i, j\} \\ \text{and } \pi_k(a_i', b_{ij}) > \pi_k(a_i, b_{ij}) \text{ for some } k \in \{i, j\} \end{array} \right\}$$

By contrast, DK's notion of $E_i = E_i^{D\&K}$ is independent of beliefs. Denote by H the set of roots of subgames. For $h \in H$, $a_i(h)$ is the strategy that prescribes the same choices as a_i, except for the choices on the path to h. For the choices on the path to h the strategy $a_i(h)$ prescribes a probability of 1. D&K use the following definition:

$$E_i^{D\&K} = \left\{ \begin{array}{c} a_i \in A_i \,\middle|\, \nexists a_i' \in A_i \text{ such that the following holds:} \\ \text{(i) } \pi_k(a_i'(h), a_j(h)) \geq \pi_k(a_i(h), a_j(h)) \text{ for all } h,\ a_j,\ k \in \{i, j\} \\ \text{(ii) } \pi_k(a_i'(h), a_j(h)) > \pi_k(a_i(h), a_j(h)) \text{ for some } h,\ a_j,\ k \in \{i, j\} \end{array} \right\}$$

In words, $E_i^{D\&K}$ excludes a strategy a_i iff there is another strategy a_i' which describes choices that lead to Pareto-superior outcomes in some parts of the game, and no Pareto-inferior or Pareto-incomparable outcome in all parts of the game, for all possible strategies of the other player.

The difference in the calculation of the equitable payoff may seem to be a subtle technicality, but it has surprisingly strong implications. To see this, we refer from now on to the most discussed application of kindness, which is of course reciprocity: This denotes an agent's inclination to be kind to someone who is kind to him, and to answer unkindness by unkind behavior. To incorporate this motivation, follow D&K and assume that player i's utility is given by his material payoff and an additional reciprocity term:

$$u_i = \pi_i(\cdot) + Y_i \cdot \kappa_{ij} \cdot \lambda_{iji},$$

where $Y_i \geq 0$ is i's "reciprocity sensitivity." Since κ_{ij} and λ_{iji} depends on i's first- and second-order beliefs, u_i is a belief-depedent utility function in the sense of so-called psychological game theory (Battigalli and Dufwenberg, 2009; Geanakoplos et al., 1989).

To get equilibrium predictions, require that all beliefs coincide with the chosen strategies and that at the root of every subgame beliefs are updated such that they are consistent with reaching this particular subgame. Furthermore, at each history the choice made is optimal given the beliefs.

---

[3] D&K's and Rabin's approaches differ also in some other ways. Most importantly, Rabin's model is defined for normal form games with constant beliefs, while D&K allow for belief updating, making their approach suitable to analyze extensive form games. But since this difference is not the topic of this note, we do not discuss it further.

[4] All beliefs are point-beliefs, assigning probability 1 to whatever is believed.

[5] This is in particular true for the normalized kindness function used in Rabin's paper – see also footnote 2.

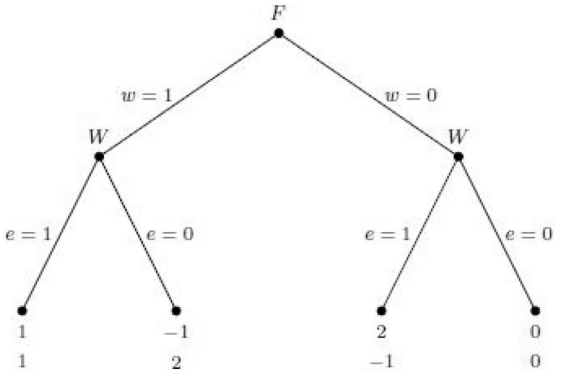
3. Three reasons to like D&K's definition

The choice of kindness/efficiency definition has become a matter of controversy. Some scholars who did applied work involving reciprocity built on D&K's model, but modified it to invoke Rabin's efficiency definition, which can matter crucially to conclusions (see e.g. Netzer and Schmutzler, 2014). On other occasions, we did applied work using D&K's definition and had referees tell us they would prefer Rabin's efficiency definition (this happened e.g. to Jang et al. (2018) see their section 5.1 for related comments). We disagree with these positions. Already D&K (in their Section 5) motivated their formulation and argued that it had certain desirable properties that Rabin's definition lacked. However, their discussion is brief, not forcefully presented, and it leaves out additional arguments that should be made. For a fuller set of arguments, read *this* section instead:

3.1. The empirical reason

Much of the literature on reciprocity is motivated with reference to gift exchange in labor markets. This classic work goes back to Akerlof (1982), Akerlof and Yellen, (1988, 1990), and is followed up by experimentalists (e.g. Fehr et al., 1997, 1993, 1998). We show that kindness a la D&K can explain gift exchange while the Rabin style definition can not.

Consider a game form where player F (the "firm") first chooses wage $w \in \{0, 1\}$ which is observed by player W (the "worker") who chooses effort $e \in \{0, 1\}$. The firms's profit equals $2e - w$ while the worker's material reward equals $2w - e$. The situation is described by Example 3, which has the structure of a sequential prisoners' dilemma:



If the players are selfish, motivated according to the given payoff numbers, there is a unique subgame perfect equilibrium: W chooses $e = 0$ regardless of w; F's best response is $w = 0$. The outcome is inefficient as both players would get higher payoffs following path $(w = 1, e = 1)$.

Akerlof and Yellen's gift exhange hypothesis is that this outcome will, in fact, materialize, and that the W's strategy will involve what has subsequently come to be called "conditional cooperation," choosing a high effort following a high wage and a low effort following a low wage. In the figure, W would choose $e = 0$ in response to $w = 0$ and $e = 1$ in response to $w = 1$. F's best response is $w = 1$. Such gift-exchange is empirically supported. That's the main message of the experimental studies we cited.

Can conditional cooperation be theoretically justified if W is motivated by reciprocity? The answer is *no* under the Rabin-style definition. To see this, note that if gift-exchange were an equilibrium then F would believe that the material payoffs following his choices $w = 1$ and $w = 0$ would be, respectively, (1,1) and (0,0). Since $(1, 1) >> (0, 0)$ the choice $w = 0$ would not belong to $E_i^{Rabin}(b_{ij})$. Only $w = 1$ would be an efficient choice, and hence *F would not be kind when choosing $w = 1$*. Therefore, W would not choose high effort in return. Gift-exchange is not viable in equilibrium.

The prediction is different with D&K's definition, according to which F's choice $w = 0$ is not inefficient. If W is sufficiently strongly motivated by reciprocity (i.e. if Y_W is high enough) then (using $E_i^{D\&K}$) the gift-exchange strategy profile is an equilibrium.

3.2. The psychological reason

Rabin's kindness definition exhibits discontinuities that we find psychologically implausible. Consider again Example 3. Assume that F believes W will respond to $w = 0$ with $e = 0$ (for sure). Assume furthermore that F assigns probability p to the prospect that W will respond to $w = 1$ with $e = 1$. Now ask: is F kind? According to the Rabin-style definition, the answer is *yes* if and only if $p \leq 1/2$; if $p > 1/2$ then choice $w = 0$ no longer sits in $E_F^{Rabin}(b_{ij})$. As p passes from below across the $p = 1/2$ hurdle, F's kindness discontinuously jumps down from a positive number to zero.

We find that counter-intuitive. It makes sense that F's kindness is decreasing in p, since after all W's payoff decreases with p, and D&K's definition captures that. But the discontinuity of F's kindness at $p = 1/2$ is questionable, because whatever F accomplishes at $p = 1/2$ is very similar to what F accomplishes at $p = 1/2 + \varepsilon$, for a small $\varepsilon > 0$.[6]

Besides discontinuity, there is a second issue involved here, namely that F's kindness would be 0 if $p > 1/2$. One may try to argue in favor of that feature on the grounds that if p exceeds $1/2$ then F should no longer be considered kind because F's beliefs lead him to think that $w = 1$ in expectation maximizes both players' material reward. That is, since choice $w = 0$ is materially Pareto-inferior (according to F's beliefs) $w = 1$ cannot be a kind choice.

We find this view non-compelling. Modelling kindness amounts to modelling human psychology, so one must ask oneself if a particular assumption is intuitively meaningful. We make this call by introspection. Does it make sense that a firm that offers its workers high wages is considered kind even if it believes that it thereby gains as well? Or, to broaden the scope, we have friends who treat us well because they seem to enjoy our friendship and they probably believe that they will benefit from remaining our friends in the future. Does this make them not kind? The answers are, in our view, respectively, *yes* and *no*.[7] We are unimpressed by the argument that a self-serving choice of $w = 1$ would not be kind.

3.3. The epistemic reason

Consider again Example 3, and the gift-exchange strategy profile involving conditional cooperation. Given equilibrium beliefs, F's choice $w = 0$ is a strategy that is not supposed to be used in that equilibrium. The strategy is moreover inefficient (in the sense of not appearing in $E_F^{Rabin}(b_{FW})$). Note that this conclusion is reached with reference to F's equilibrium beliefs.

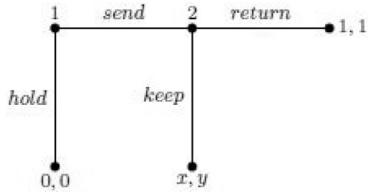
We find it is questionable to evaluate the motives that underlie a non-equilibrium choice with the maintained assumption that F has equilibrium beliefs. If F chooses $w = 0$ he demonstrably is *not* playing along with the gift-exchange equilibrium. This undermines the argument that if he chooses $w = 0$ he must be doing something that he views as inefficient.

The issue is analogous to ones that come up in epistemic game theory, concerning a difficulty with justifying backward induction (BI). The critique was first articulated by Kaushik Basu and Phil Reny in the mid-1980s. Recall the issue: Suppose that player i deviates from the BI path, that j is then asked to move, and that i will later move again. BI, implicitly, calls for j to assume that i will conform with BI in the future. Maintaining that belief is most awkward, however, as j has already seen evidence that i is, in fact, not a backward-inducter. If j therefore considers it possible that i may not conform with BI going forwards, he may have reason to deviate from BI himself. And if i then figures out that j may reason that way, then he may have an incentive to deviate from the BI path to start with![8]

The corresponding issue in our context concerns how to judge the motives of a player who deviates from the equilibrium path, as regards whether or not he would be attempting to lower the payoffs of all players. Even if the player does not get to move again (as was essential in the BI-case), this matters as regards the determination of what he might have considered inefficient *had he chosen it*. D&K's definition identifies a set of strategies (the inefficient ones) for which no story can be told that explains the behavior unless that story involves "waste" at some history. In our view, this is a set with a meaningful interpretation, which helps one to think about and interpret the zero-kindness-threshold.

4. I&S' critique and a new E_i-definition

I&S argue that kindness-based reciprocity models cannot explain beneficial trust in simple trust games. Consider the following: Player 1 has to decide between *hold* and *send*. If he decides to *hold*, the game ends with material payoffs of (0,0), while in case of *send* player 2 has to make a choice between *keep* and *return* with respective material payoffs of (1,1) and (x, y):



[6] One can furthermore construct examples where this sort of discontinuity precludes equilibrium existence; see D&K (p. 289). The main problem with discontinuity, as we see it, however, is not with non-existence *per se*, but with the underlying psychology being implausible (which then happens to cause non-existence). Relatedly, one can get around the existence issue by 'smoothening' utilities in various way; see Rabin's appendix for relevant pointers. However, what is psychologically implausible is not just discontinuity but that the zero-threshold-kindness varies "quickly" with p.

[7] This view is also supported by research in cognitive and evolutionary psychology. According to e.g. Fiske et al. (2007) and Schulz (2016) inferences about the level of altruism (in our terminology about the kindness) of other human beings are cognitively less costly and faster than inferences about egoistic motives of others.

[8] The power of this argument is perhaps seen most starkly in centipede games. For relevant references and further commentary, see (Basu, 1988; Reny, 1988, 1993), and Asheim and Dufwenberg (2003). Petitt and Sugden (1989) make comparable arguments for finitely repeated prisoners' dilemma games.

I&S first consider a trust game they label G_1 with $(x, y) = (-1, 3)$, which may be seen as a simplified version of the gift-exchange game described above. For the same reason as in the gift-exchange game, (*send, return*) cannot be an equilibrium for kindness-based models if one uses E_i^{Rabin} for the calculation of the equitable payoff. I&S label this the "Paradox of Trust."

The paradox disappears if one uses D&K's approach, as I&S acknowledge. Using $E_i^{D\&K}$, (*send, return*) is a reciprocity equilibrium if Y_2 is high enough. But I&S go on to argue that $E_i^{D\&K}$ does not lead to plausible kindness evaluations in other games, and to prove their point they consider games G_2 where $(x, y) = (1/2, -1/2)$ and G_3 where $(x, y) = (1/2, 1/2)$. They make the following argument: If for a given belief *send* is regarded as kind in G_2, then it should be regarded as kind in G_3 for the same belief, since the only difference between the games is that 2's material payoff of (*send, keep*) is larger in G_3 than in G_2. In G_3 the strategy *hold* $\notin E_1^{D\&K}$, implying that *send* would carry 0 kindness. On the other hand, in G_2 the strategy *hold* $\in E_1^{D\&K}$, implying that now the strategy *send* is regarded as kind for the belief that player 2 plays *return*.

I&S intriguing observation points to a flaw in the definition of $E_i^{D\&K}$.. The problem can be overcome by instead using the following new definition of efficient strategies (which is still independent of any beliefs):

$$E_i^{new} = \left\{ a_i \in A_i \middle| \nexists a_i' \in A_i \text{ such that the following holds:} \begin{array}{l} \text{(i) } \pi_i(a_i'(h), a_j(h)) \geq \pi_i(a_i(h), a_j(h)) \text{ for all } h,\ a_j, \\ \text{(ii) } \pi_i(a_i'(h), a_j(h)) > \pi_i(a_i(h), a_j(h)) \text{ for some } h,\ a_j. \end{array} \right\}$$

This definition is similar to that of the original $E_i^{D\&K}$, except that the "efficiency" of strategies of i is only checked with respect to i's material payoffs, and not with respect to that of the other player. To see the intuition behind this definition, recall that the purpose of introducing E_i is to determine the equitable payoff such that it is not influenced by i engaging in "waste" (see also the commentary at the end of Section 3.3). Note that in general $E_i^{new} \subseteq E_i^{D\&K}$, and recall (from section 2) that E_i is only relevant for the calculation of $\min_{a_i \in E_i} \pi_j(a_i, b_{ij})$, but not for $\max_{a_i \in A_i} \pi_j(a_i, b_{ij})$. Elements of $A_i \setminus E_i^{D\&K}$ obviously qualify as wasteful. But so do elements of $E_i^{D\&K} \setminus E_i^{new}$. Too see this, suppose that $\tilde{a}_i \in E_i^{D\&K} \setminus E_i^{new}$ and that the inclusion of $\tilde{a}_i$ in E_i makes $\min_{a_i \in E_i} \pi_j(a_i, b_{ij})$ go lower. Reflect on the impact of $\tilde{a}_i$. Given i's beliefs it both reduces j's material payoff and it involves material loss for i (in the sense of (i) and (ii) in the definition of E_i^{new}).[9] Hence, D&K's original notion of inefficient strategies seems slightly (and subtly) too conservative.

This new efficiency notion takes care of the problem identified by I&S. For G_2 it holds that $E_1^{new} = \{send\}$ is a strict subset of $E_1^{D\&K} = \{hold, send\}$. For any values of Y_1 and Y_2, using E_1^{new}, the strategy profile (*send, return*) is a reciprocity equilibrium in G_2 as well as G_3 and the kindness associated with choice *send* is the same in either game (namely, $= 0$).[10] It can be easily checked that for high enough Y_2, (*send, return*) is also a reciprocity equilibrium of G_1, and $(w = 1, e = 1)$ is a reciprocity equilibrium in Example 3. Furthermore, all the arguments in favor of D&K's original notion of efficient strategies (see section 3) apply also to this adapted definition.

In practice, the difference between the two efficiency notions seems limited. Whether one relies on E_i^{new} or $E_i^{D\&K}$ doesn't matter in most games we encountered, the exceptions being G_2 and (for n-player games) Hahn's (2009) voting game (see footnote 9). Moreover, an analog of D&K's proof for existence of equilibrium applies using E_i^{new} instead of $E_i^{D\&K}$.

5. Final words

We started out giving reasons why kindness may influence economic outcomes. Our subsequent discussions was focused on reciprocity, but we think that the concept we ended up advocating (a slightly revised version of Dufwenberg and Kirchsteiger, 2004) is relevant more generally. We close out quoting the war cry of *Ben's Bells*, a non-profit organization which creates and distributes handmade bells:[11] "Be kind."

References

Akerlof, G., 1982. Labour contracts as a partial gift exchange. Q. J. Econ. 97, 543–569.
Akerlof, G., Yellen, J., 1988. Fairness and unemployment. Am. Econ. Rev. 81, 1096–1135. (Papers & Proceedings).
Akerlof, G., Yellen, J., 1990. The fair-wage effort hypothesis and unemployment. Q. J. Econ. 195, 255–284.
Aldashev, G., Sebald, A., Kirchsteiger, G., 2017. Assignment procedure biases in randomized policy experiments. Economic Journal 127, 873–895.
Asheim, G., Dufwenberg, M., 2003. Deductive reasoning in extensive form games. Econ. J. 113, 305–325.
Basu, K., 1988. Strategic irrationality in extensive games. Math. Soc. Sci. 15, 247–260.
Battigalli, P., Dufwenberg, M., 2009. Dynamic psychological games. J. Econ. Theory 144, 1–35.
Bierbrauer, F., Netzer, N., 2016. Mechanism design and intentions. J. Econ. Theory 163, 557–603.
Bierbrauer, F., Ockenfels, A., Pollak, A., Ruckert, D., 2017. Robust mechanism design and social preferences. J. Public Econ. 149, 59–80.
Conconi, P., DeRemer, D., Kirchsteiger, G., Trimarchi, L., Zanardi, M., 2017. Suspiciously timed trade disputes. J. Int. Econ. 105, 57–75.
van Damme, E., Binmore, K., Roth, A., Samuelson, L., Winter, E., Bolton, G., Ockenfels, A., Dufwenberg, M., Kirchsteiger, G., Gneezy, U., Kocher, M., Sutter, M., Sanfey, A., Kliemt, H., Selten, R., Nagel, R., Azar, O., 2014. How werner guth's ultimatum game shaped our understanding of social behavior. J. Econ. Behav. Organ. 108, 292–318.

[9] The notion of E_i^{new} can easily generalized for n-player games by replacing index j by $-i$ in the definition above, with $-i$ meaning "all players but i." The same intuition as for the two-player games applies. We note that such a definition overcomes a concern that Hahn (2009, p. 469) highlighted for an n-player voting game, which lead him to suggest a modified equitable-payoff definition for his context.

[10] In G_2, for high enough Y_2 there is another equilibrium where the players are unkind to each other: (*send, keep*).

[11] The mission of this Tucson-based organization is to "inspire, educate, and motivate people to realize the impact of intentional kindness, and to empower individuals to act according to that awareness, thereby strengthening ourselves, our relationships and our communities."

Dufwenberg, M., Gächter, S., Hennig-Schmidt, H., 2011. The framing of games & the psychology of play. Games & Economic Behavior 73, 459–478.
Dufwenberg, M., Kirchsteiger, G., 2000. Reciprocity & wage undercutting. Eur. Econ. Rev. 44, 1069–1078.
Dufwenberg, M., Kirchsteiger, G., 2004. A theory of sequential reciprocity. Games Econ. Behav. 47, 268–298.
Dufwenberg, M., Patel, A., 2017. Reciprocity networks & the participation problem. Games Econ. Behav. 101, 260–272.
Dufwenberg, M., Rietzke, D., 2016. Banking on reciprocity: deposit insurance & insolvency. Mimeo..
Dufwenberg, M., Smith, A., Essen, M.V., 2013. Hold-up: with a vengeance. Econ. Inq. 51, 896–908.
Falk, A., Fischbacher, U., 2006. A theory of reciprocity. Games Econ. Behav. 54, 293–315.
Fehr, E., Gächter, S., Kirchsteiger, G., 1997. Reciprocity as a contract enforcement device: experimental evidence. Econometrica 65, 833–860.
Fehr, E., Kirchsteiger, G., Riedl, A., 1993. Does fairness prevent market clearing? an experimental investigation. Q. J. Econ. 108, 437–460.
Fehr, E., Kirchsteiger, G., Riedl, A., 1998. Gift exchange and reciprocity in competitive experimental markets. Eur Econ Rev 42, 1–34.
Fiske, S., Cuddy, A., Glick, P., 2007. Universal dimensions of social cognition: warmth and competence. Trends Cogn. Sci. (Regul. Ed.) 11, 77–83.
Geanakoplos, J., Pearce, D., Stacchetti, E., 1989. Psychological games and sequential rationality. Games Econ. Behav. 1, 60–79.
Hahn, V., 2009. Reciprocity and voting. Games Econ. Behav. 67, 467–480.
Isoni, A., Sugden, R., 2018. Reciprocity and the paradox of trust in psychological game theory. J. Econ. Behav. Organ.. Forthcoming
Jang, D., Patel, A., Dufwenberg, M., 2018. Agreements with reciprocity: co-financing and MOUs. Games Econ. Behav. 111, 85–99.
Le Quement, M., Patel, A., 2018. Cheap talk as gift exchange. mimeo.
Netzer, N., Schmutzler, A., 2014. Explaining gift-exchange – the limits of good intentions. J. Eur. Econ. Assoc. 12, 1586–1616.
Netzer, N., Volk, A., 2014. Intentions & ex-post implementation. mimeo.
Nyborg, K., 2018. Reciprocal climate negotiators. J. Environ. Econ. Manage.. Forthcoming
Petitt, P., Sugden, R., 1989. The backward induction paradox. Journal of Philosophy 4, 169–182.
Rabin, M., 1993. Incorporating fairness into game theory and economics. American Economic Review 83, 1281–1302.
Reny, P., 1988. Rationality, Common Knowledge and the Theory of Games. In: Phd dissertation, chapter 1, department of economics. Princeton University
Reny, P., 1993. Common belief and the theory of games with perfect information. J. Econ. Theory 59, 257–274.
Schulz, A., 2016. Altruism, egoism, or neither: a cognitive-efficiency-based evolutionary biological perspective on helping behavior. Stud. Hist. Philos. Biol. Biomed. Sci. 56, 15–23.
Sebald, A., 2010. Attribution and reciprocity. Games Econ. Behav. 68, 339–352.
Segal, U., Sobel, J., 2007. Tit for tat: foundations of preferences for reciprocity in strategic settings. J. Econ. Theory 136, 197–216.
Sobel, J., 2005. Interdependent preferences and reciprocity. J. Econ. Lit. 43, 396–440.