Incidental ostracism emerges from simple learning mechanisms

Lindström, Björn ; Tobler, Philippe N

DOI: https://doi.org/10.1038/s41562-018-0355-y

Incidental ostracism emerges from simple learning mechanisms

Björn Lindström[1, 2] & Philippe N. Tobler[1]

1. Laboratory for Social and Neural Systems Research, Department of Economics, University of Zurich, Zurich, Switzerland

2. Section for Psychology, Department of Clinical Neuroscience, Karolinska Institute, Stockholm, Sweden

Abstract: Ostracism, or social exclusion, is widespread and associated with a range of detrimental psychological and social outcomes. Ostracism is typically explained as instrumental punishment of free-riders or deviants. However, this instrumental account fails to explain many features of real-world ostracism, including its prevalence. We hypothesized that ostracism can emerge incidentally (non-instrumentally) when people choose partners in social interactions, and that this process is driven by simple learning mechanisms. We tested this hypothesis in four experiments (n = 456) with economic games on dynamic social networks. Contrary to the instrumental account of ostracism, we find that the targets of ostracism are not primarily free-riders. Instead, incidental initial variability in choosing partners for social interactions predicts later ostracism better than the instrumental account. Using computational modeling, we show that simple reinforcement learning (RL) mechanisms explain the incidental emergence of ostracism, and that they do so better than a formalization of the instrumental account. Finally, we leveraged these RL mechanisms to experimentally reduce incidental ostracism. Our results demonstrate that ostracism is more incidental than previously assumed and can arise from basic forms of learning. They also show that the same mechanisms that result in incidental ostracism can help to reduce its emergence.

Ostracism, or the exclusion of an individual from social interaction, is a persistent phenomenon in human groups[1]. Ostracism is distinct from the other major type of social exclusion, social rejection[2]. While ostracism consists of simply neglecting or avoiding another individual[2], and is subjectively characterized by the feeling of being ignored[3], social rejection is explicit, purposeful and often aggressive in nature[2]. However, despite its relative subtlety, ostracism can have more severe consequences than explicit social rejection and even bullying for both the mental and physical health of the excluded individual[1,4]. Indeed, a large body of evidence demonstrates that being ostracized is highly aversive, resulting in threatened fundamental psychological needs of belonging, self-esteem, meaningful existence, and sense of control[1], reduced work satisfaction and increased staff turnover rates[3], increased physiological stress responses[5], and activation of brain regions that overlap with those processing physical pain[6].

The detrimental consequences of ostracism are typically attributed[1,7] to its putative primary evolutionary function as a low cost solution to the free-rider problem[8–13]. In line with this view, theoretical accounts of ostracism commonly conceptualize it as a motivated instrument to punish or discipline free-riders and other annoying deviants[7,14,15], such as aggressive upstarts[16] or individuals with a contagious disease[9]. Experimental studies support this *instrumental* account of ostracism, by showing that ostracism, especially if it is coordinated, can promote cooperation[17–21]. Similarly, social psychology research has shown that ostracism is used instrumentally to punish burdensome individuals (e.g., individuals who are dispositionally disagreeable[22], which is predictive of free-riding[23], or individuals who are bad co-players in computerized games[24–26]).

However, there are several characteristics of real world ostracism that are not well accounted for by the instrumental view. First, the victims of real world ostracism often appear to be selected randomly[27]. Furthermore, the incidence of ostracism reported in samples from the general population is very high (e.g., daily ostracizing events)[28], and episodes where the

ostracizers (sources) are oblivious of the ostracized (target), or where the reason for ostracism is unclear, are more common than episodes of instrumentally motivated ostracism[28]. For example, a recent study of workplace ostracism demonstrated that an individual's perception of being ostracized was unrelated to others' intention to ostracize[29]. Moreover, ostracism often takes place in dyadic interactions[28,30], rather than in coordinated group actions that have been the focus of research on instrumental ostracism and cooperation[18–20]. These findings suggest that a substantial degree of experienced ostracism might be *incidental* rather than instrumental, raising the possibility that the reasons for the high prevalence of ostracism among humans are not fully accounted for by the instrumental view.

Here, we propose that ostracism among humans can be caused incidentally by self-organization of social interaction partners who intend to maximize rewards and avoid punishment, *without* instrumental motivation for ostracism. This form of ostracism is an *emergent* phenomenon, a macro-level pattern that is not reducible to beliefs and desires of individuals to ostracize[31], which entails a radically different explanation from the instrumental account. The instrumental and emergence accounts of ostracism are complementary rather than mutually exclusive, as it is clear that instrumental ostracism occurs in the real world[30] and the two forms of ostracism can co-exist in principle. Importantly, although the possibility that experienced ostracism might be incidental has been proposed previously[3,32], the scope of, and the actual mechanisms causing, incidental ostracism are not well understood. We tested, and corroborated, the emergence account of ostracism using real-world social networks, a series of behavioral experiments, and computational analyses.

A growing empirical literature on dynamic social networks provides clues on how ostracism could emerge incidentally. When individuals choose partners to interact with, they form and sever network links with others, which can be beneficial for cooperation[33–36]. Cooperation is facilitated because partner choice establishes relationships (links) between

individuals contingent on past behavior (e.g., being cooperative). Crucially, this implies that as people select partners, others are excluded from that interaction. This will be the case even if individuals have no intention of punishing free-riders, but simply seek out payoff maximizing interactions. Thus, ostracism might be a natural consequence, or by-product, of partner choice in social networks, a relationship that appears to have been largely overlooked (but see a discussion comment in ref. 34, p. 6). Furthermore, theoretical work on the formation of social networks through partner choice has shown that associative reinforcement learning (RL), where agents repeat rewarded actions and avoid punished actions, can result in considerable social network structure, where some agents are less popular than others, even in simple games[37].

Against this background, we hypothesized that ostracism can emerge incidentally if people make choices about interaction partners based on what they learnt from previous interactions. Incidental ostracism would thereby be a side effect of the same processes that lead to reciprocity in social networks. Distinct from the instrumental account, which predicts that people instrumentally ostracize poor interaction partners, such as free-riders[11,17], our emergence account predicts ostracism (some people are excluded from interactions) also in the *absence* of free-riding or deviant behavior. To illustrate, if students on the first day of term randomly select a possible friend in the class or dormitory, and continue to interact with the same person if the experience was positive, some students might become, and stay, ostracized purely by chance (i.e., without free-riding). Crucially, this would be evident even if the ostracized students objectively were as nice as the popular ones, and no motivation for ostracism existed in the group. This is an example of *path dependence*[38], which entails that early random outcomes probabilistically constrain the future. Path dependence in ostracism is a key prediction of the emergence account: the learning process should generate increasing returns for repeated interactions with the same social partners, even if these partners were initially randomly

selected, as any beneficial interaction (e.g., mutual cooperation) with a specific partner will increase the expected value of interacting with this partner in the future[38].

To test the emergence account, we used a combination of real-world social networks, behavioral lab experiments, and computational agent-based modeling (ABM). Our experimental model of social interaction (drawing inspiration from theoretical work[37]) had groups of six participants (players) playing an iterated Prisoner's Dilemma (Experiment 1 and Experiments 3-4) or coordination (Experiment 2) game for monetary payoffs (Figure 1). Each player selected one partner in each experimental period (i.e., established a directed link) and could be selected by multiple other players (max 5) as partner in each period. We tested if the resulting evolving social network would be path dependent, so that early variability in partner choices would lead some players to become popular and others to be largely left out (incidental ostracism). In Experiment 2 (coordination game), we investigated the incidental emergence of ostracism in a game without free-riding, in order to experimentally rule out instrumental punishment of free-riders as an explanation for ostracism. These two experiments confirmed the prediction that ostracism can be path dependent and emerge incidentally. We formalized the emergence account as an agent-based RL model, which provided a superior account of the data relative to a formalization of the instrumental account, and showed that RL mechanisms are sufficient for generating incidental ostracism. Finally (Experiments 3-4), guided by the mechanistic RL model, we devised an experimental intervention that reduced the emergence of incidental ostracism.

Results

Real-world social networks confirm a central prediction of the emergence account

Do real-world social networks exhibit the basic characteristics predicted by the incidental emergence account of ostracism? To answer this question, we first analyzed four pre-existing longitudinal data sets, involving the formation of friendships in school classes and among college freshmen (see Supplementary Methods for details).

People differed markedly in their popularity in all four social networks. The *Gini* coefficient, which quantifies the degree of inequality (0 = no inequality, 1 = maximum inequality) in the distribution of a resource (here, popularity), ranged between .32 and .39. This shows that while some people were highly popular, others were ostracized. A core prediction of the emergence account is that popularity (and hence ostracism) will be *path dependent*. Path dependence entails that a system's early states probabilistically constrain its future states[38]. In keeping with path dependence, early popularity (the first mapping of the social network) predicted late popularity (the last mapping of the social network) in all four datasets (Figure 2, largest p – value = .008).

These data are in line with a central prediction of the emergence account – that there is a positive relationship between early and late popularity – and suggest that ostracism might emerge in an incidental manner in the real world. However, as the data sets did not track the behavior of the respondents (e.g., free-riding), we cannot rule out instrumental ostracism as an explanation for the relationship between early and late popularity. Similarly, we cannot rule out the contribution of stable traits such as attractiveness (e.g., students that are more attractive might be more popular both at early and late timepoints). For these reasons, we next conducted a series of lab experiments where we carefully contrast predictions from the emergence account and the instrumental account, by testing how path dependence and free-riding contribute to ostracism. Note that participants in the experiments were completely anonymous (see Methods). Therefore, individual differences in stable traits (e.g., physical attractiveness) can be ruled out as alternative explanations for ostracism.

Objective ostracism arises during task and predicts subjective ostracism

In our first experiment, anonymous participants (n = 186) repeatedly selected partners and actions during iterated Prisoner's Dilemma games in a dynamic social network with fixed size (see Figure 1 and Supplementary Figure 1). The PD is the quintessential example of a social dilemma, and thereby ideal for contrasting the instrumental account with the emergence account of ostracism.

In the experiments, we defined objective ostracism as the difference between the network in-degree (i.e., popularity) expected if each player was selected as partner once per round (i.e., 35, the absence of ostracism), and the actual in-degree (the total number of times that a participant was chosen as a partner by the other 5 group members). Positive differences thereby indicate players who were selected less than once per period on average, and negative differences indicate players who were selected more than once per period on average. Objective ostracism was thus by definition relative in our experiments; some people were ostracized to the degree that others were popular (i.e., ostracism is the inverse of popularity).

We used the objective ostracism measure first to characterize the distribution of ostracism in our experimental networks. Consistent with the real social networks analyzed above, the Gini coefficient was ~.32 (bootstrapped 95% CI [.29, .36]), indicating considerable variability in objective ostracism. Furthermore, the distribution of objective ostracism was markedly different from what would have been the case if participants randomly selected a partner in each period (bootstrapped Kolmogorov-Smirnov [K-S] test: $D = 0.34$, $p < .0001$, see Figure 3a), which corresponds to a Gini coefficient of ~.087 (95% CI [.086, .089]). These results reflect considerable variance in objective ostracism, which is required to compare the different accounts of ostracism.

Next, we confirmed that objective ostracism was tightly related to the subjective experience of ostracism. We captured subjective ostracism as the mean of the four fundamental psychological needs measured by the *need-threat* scale[39] (Cronbach's α = 0.91), a standard measure of subjective ostracism, completed after the experiment. We found a strong predictive relationship such that higher objective ostracism was associated with stronger subjective ostracism (see Figure 3b). This relationship remained significant when controlling for total payoff, both with the composite measure and with the individual subscales (see Supplementary Table 2). These results demonstrate that our measure of objective ostracism was strongly associated with the subjective experience of ostracism and thereby provide construct validity for our experimental approach[2].

Objective ostracism is path dependent as predicted by the emergence account

Next, we contrasted explanations from the instrumental and emergence accounts for the occurrence of objective ostracism in Experiment 1. The instrumental account holds that free-riding behavior should strongly predict ostracism whereas the emergence account predicts that objective ostracism should be path dependent. Path dependence means that early random variability (i.e., variability that is not explained by differences in early behavior) in who is selected as a partner should have a strong effect on later objective ostracism. In the emergence account, path dependence explains why observed objective ostracism can be incidental rather than instrumental. We found above that real social networks exhibit a correlation between early and late popularity, suggestive of path dependence (see *"Real-world social networks confirm a core prediction of the emergence account"*). Crucially, here we assessed if this relationship still holds when free-riding behavior is accounted for, which is necessary for establishing path dependence as an explanation for incidental ostracism[38]. In other words, path-dependence entails that ostracism should occur even in the absence of free-riding (see Supplementary Note

1 for additional analyses showing that partner choice was sensitive to free-riding, an effect predicted by both accounts), which is not predicted by the instrumental account.

We tested these distinct predictions of the emergence and the instrumental account by assessing how free-riding and path dependence explained later ostracism. Specifically, we used objective ostracism and free-riding behavior, operationalized as the proportion of Defect actions as initiator and responder, during the initial five periods as predictors of objective ostracism in the *final* five period (objective ostracism in the five first periods was positively correlated with free-riding both as initiator and responder [*r*s = .45 and .39]. Variance inflation factor [VIF] analyses demonstrated that these correlations did not cause problematic regression collinearity: max VIF = 2.94. Moreover, the findings are robust to alternative, non-linear, model specifications, see Supplementary Tables 3-4). As predicted by the emergence account, early objective ostracism strongly predicted objective ostracism at the end of the experiment when controlling for free-riding (β = 1.2, SE = 0.24, t = 5.1, $p < .0001$, Figure 4a-b). In contrast, early free-riding did not reliably predict final objective ostracism, neither as initiator nor as responder (smallest p = .09, see Figure 4b and Supplementary Note 1 for further analyses). In a control analysis, we tested whether path dependent ostracism was in place also earlier during the experimental session and found that it was. Specifically, we could predict ostracism in periods 11-15 (rather than 31-35) from initial ostracism (β = 1.67, SE = 0.21, t = 7.82, $p <$. 00001), controlling for free- riding. Thus, our results are robust to the exact number of social interaction periods. Together, path dependence, rather than free-riding, was the best predictor of ostracism, as anticipated by the emergence account.

As an additional test of path dependence, we conducted the same analysis for the subset of participants (n = 64) who never defected as initiator (behavior as initiator is a more accurate measure of free-riding tendency than as responder, as the latter also includes responses to other players' defections), and found again a strong path dependent relationship (β = 1.79,

SE = 0.44, t = 3.9, p < .0001). This analysis shows, in analogy to our introductory example about freshman students, that despite having objectively equivalent, or even better, quality as partners, some players initially became, and subsequently stayed, ostracized. Notably, objective ostracism was tightly related to the subjective experience of ostracism also in this sub-group (β = 0.28, SE = 0.08, t = 3.58, p < .0001), which demonstrates the detrimental subjective consequences of incidental ostracism.

Together, these results demonstrate path dependence for objective ostracism; early variability in objective ostracism best explained later objective ostracism, which indicates that ostracism, as predicted, can be to a large degree incidental. The emergence account, but not the instrumental account, fully predicts these results. Thus, ostracism is more incidental than previously assumed.

Path dependent ostracism in a pure coordination game without free-riding

In Experiment 2 (n = 90), we tested whether ostracism would occur and exhibit path dependence in social interactions without the possibility of free-riding. Specifically, participants interacted in a pure coordination game (often referred to as "choosing sides"). In this game, in contrast to the Prisoner's Dilemma game, there is no tension between cooperation and selfishness. Instead, to receive reward, the participants had to coordinate on one of two possible actions. This game is often used to exemplify the logic of social norms, such as left- or right-hand side traffic – it doesn't matter which side you drive on, as long as it's on the same side as everyone else. Accordingly, the instrumental account of ostracism makes no predictions about this type of game, as there is no free-riding to punish. In contrast, our emergence account predicts that ostracism should emerge in much the same way as in the Prisoner's Dilemma situation of Experiment 1.

In line with this prediction, we found that neither the Gini coefficient (~ .37, 95 % CI for difference to Experiment 1 [-0.02, .11]), nor the distribution (see Supplementary Figure 2) of objective ostracism in Experiment 2 were significantly different from those of Experiment 1 (K-S test: D = .06, p = .91). Furthermore, the degree of objective ostracism again predicted subjective ostracism (Supplementary Figure 4) (controlling for total payoff) (β = 0.28, SE = 0.12, t = 2.3, p = .02. Removing one participant with a standardized residual exceeding 2.5 SD: β = 0.32, SE = 0.11, t = 2.86, p = .005). Thus, ostracism was just as prevalent when free-riding was not possible (Experiment 2) as when it was possible (Experiment 1).

Crucially, ostracism was path dependent also in Experiment 2 (Figure 4c): the degree of objective ostracism in the first five periods strongly predicted objective ostracism in the final five periods (β = 2.82, SE = 0.45, t = 6.63, p < .0001). In analogy to Experiment 1, we controlled for coordination failures as both initiator and responder (smallest p = .18, see Figure 4d). Limiting the analysis to participants without coordination failures showed the same pattern (no failures as responder, N = 46: β = 2.57, SE = 0.57, t = 4.5, p < .0001; no failures as initiator, N = 15: β = 1.39, SE = 0.42, t = 3.29, p = 0.004). Thus, Experiment 2 demonstrates path dependent, incidental ostracism in social interactions where free-riding is not possible. The instrumental account cannot readily explain these results, which in contrast are directly predicted by the emergence account of incidental ostracism.

An Agent-Based Model demonstrates that reinforcement learning mechanisms are sufficient for the incidental emergence of ostracism

Our emergence account postulates that ostracism arises incidentally from the interaction of individuals who select their partners based on simple associative RL mechanisms. To formally test this explanation for the incidental ostracism observed in Experiments 1 and 2, we developed

a generative, agent-based simulation model where each agent´s behavior was controlled by RL mechanisms (the emergence model)[40,41]. We contrasted the emergence model against a formalization of the instrumental account (the instrumental model). Both models were devoid of parameters that directly regulated group-level characteristics; thus all results arise from the interaction of independent, simple, but psychologically plausible, agents[41]. We first describe the two models. Next, we compare them formally and find that the emergence model explains the individual-level data substantially better than the instrumental model. We then simulate the models to show that the emergence, but not the instrumental, model is sufficient for explaining the path dependent emergence of incidental ostracism. Finally, we find that the same model generalizes out-of-sample from the Prisoner's Dilemma (Experiment 1) to the coordination game (Experiment 2).

Emergence model. The agents selected partners and actions (exactly as in the experiments), and updated the values associated with both through basic RL (see Supplementary Methods for details). The basis for each agent was the standard RL (Rescorla-Wagner) learning rule, which specifies how the prediction error – the difference between the experienced outcome and the expected outcome – drives learning[42]. Crucially, the agents did not instrumentally punish defectors using ostracism, but cared only about their own payoffs from each interaction (as both responder and initiator) (see Supplementary Figure 2-3 for a description of how the model's internal dynamics and parameters generate path dependent ostracism).

Instrumental model. According to the instrumental account, ostracism is a way to punish free-riders. Thus, in analogy to strategies for the standard iterated Prisoner's Dilemma game, the instrumental account is in essence a "grim" strategy[43], where a defecting agent will be deterministically and indefinitely ostracized (not chosen as partner). This basic model was generalized to the full space of more flexible and forgiving strategies, by parameterizing ostracism (as well as Prisoner's Dilemma action selection) as probabilistic rather than

deterministic (see Supplementary Methods for details). Thus, according to the instrumental model, and in contrast to the emergence model, agents instrumentally punish defectors using ostracism, without regard for their own outcomes. The instrumental model thereby corresponds to a behavioral heuristic, conceptually similar to strategies for the repeated Prisoner's Dilemma game (e.g., Tit for Tat).

Model comparison. We fitted the two models, which had the same number of free parameters (i.e., 3), to the individual-level trial-by-trial choice data in Experiment 1 (n = 186). Model fits served to contrast the relative explanatory power of the two accounts (see Supplementary Methods for details). Model comparison strongly favored the emergence model: the choices of 168 out of 186 participants were better explained by the emergence model than the instrumental model with Δ BIC > 2. 164 participants had Δ BIC > 10 in favor of the emergence model. Using AIC instead of BIC gives identical results. In contrast, the choices of only four participants were better explained by the instrumental account (see Figure 5a). The exceedance probability for the emergence model, which denotes the probability that a model is the most common in the population, was 1 (see Figure 5b). These results provide strong support for the emergence model outperforming the instrumental model as an account of the experimental data (see Supplementary Table 9 for an analysis how estimated model parameters relate to empirical ostracism, and Supplementary Note 2 for a verification that learning, rather than choice perseverance or inertia, is needed for explaining our results).

Model validation. Evidence that one model explains the data better than an alternative model is only the first step in model comparison, as a better fit does not guarantee that the winning model actually can reproduce the effects of interest[44]. We used two additional approaches to validate the emergence model: tests of *generative performance* and *model generalizability*[44]

First, we tested the capability of the emergence model to reproduce the effects of interest, i.e., its generative performance. To this end, we simulated the emergence model (with parameter values fixed to the median of the estimated individual parameters) and submitted the simulated data to the same statistical test as our empirical results. Thus, we tested if the model, with empirical parameter values, reproduced path dependent ostracism (note that this is a more stringent criterion than only evaluating the fitted values[44,45]). This was indeed the case: as shown in Figure 6 (a-b), the emergence model faithfully generated path dependence. In contrast, the instrumental model, based on empirical parameter values, did not (Figure 6, c-d). This shows, in concert with model comparison, that the emergence model but not the instrumental model easily captures the observed path dependent ostracism. In the Supplementary Note 2 and Supplementary Figure 9, we in addition show that the emergence model, but not the instrumental model, readily generates correlations between early and late popularity of the same magnitude as observed in the real-world social networks (Figure 2).

Second, we assessed the generalizability of the emergence model by generating out-of-sample predictions (based on parameter values calibrated for Experiment 1) for Experiment 2. Note that Experiments 1 and 2 used different paradigms (Prisoner's Dilemma and coordination game, respectively), which naturally makes out-of-sample prediction difficult. Despite this difficulty, the model predictions provided a good match to the data (Figure 7), which verifies the generalizability of the emergence model (we did not conduct the same analysis for the instrumental model; as free-riding was not possible in the coordination game, the instrumental model does not make any predictions for choice behavior in Experiment 2).

Together, these results demonstrate that a simple RL model, without instrumental ostracism, is sufficient for explaining path dependent incidental ostracism. In contrast, the formalization of the instrumental account was unable to do so. Furthermore, the results show that the same RL mechanisms are valid across game types (Prisoner's dilemma vs.

coordination), which constitutes strong support for the learned basis of ostracism proposed by the emergence account.

Causal manipulation of path dependence reduces objective and subjective ostracism

The model validation clearly supports our account that the incidental emergence of ostracism is caused by basic RL mechanisms. To provide even stronger evidence for this view, we next utilized these mechanisms to causally manipulate path dependence, by setting the group on a specific path. We conducted two experiments that varied the strength of this manipulation. In Experiment 3 (n = 90), each player was paired up with another player in the first period and then allowed to freely select partners for the remaining 35 periods (exactly as in Experiment 1). Similarly, in Experiment 4 (n = 90) we initially paired each player up with another player but this time for five initial periods, and then allowed them to freely select partners for the remaining 35 periods. In both Experiments 3 and 4, the social interaction during paired and free choice trials occurred in the form of the Prisoner's Dilemma game (as in Experiment 1; Experiment 1 was also comparable to Experiments 3 and 4 in the distribution of participant age, gender, and the degree of free-riding, see Supplementary Note 1). The model, with parameters estimated from Experiment 1, predicted that this manipulation of path dependence would reduce objective ostracism, if free-riding during the paired periods is accounted for (which should increase ostracism according to both incidental and instrumental accounts: predicted β = 3.7) (Figure 8). In other words, the emergence model predicts that path dependence can be leveraged as an intervention against the emergence of incidental ostracism, by pairing up players.

The proportion of periods where the players selected the same partner as they interacted with in the first period (Experiment 1: 0.17, Experiment 3: 0.39, Experiment 4: 0.5),

was considerably higher in Experiments 3 and 4 than in Experiment 1 (Welch tests: t(101.34) = 5.47, $p < .0001$, t(98.76) = 7.48, $p < .0001$). The tendency to stick with the same partner was furthermore slightly stronger in Experiment 4 than Experiment 3 (t(175.6) = 1.96, $p = .049$). These results show that the manipulation was successful in setting the group on a particular path (see Supplementary Figure 9 and the Supplementary Note 1 for additional analyses of path dependence in Experiments 3 and 4). Next, we tested the model prediction that manipulating path dependence would reduce objective ostracism (Figure 8a). Confirming this prediction, objective ostracism, averaged across the entire experiment, was significantly lower in Experiment 3 (β = -6.79, SE = 2.76, t = -2.46, $p = .015$) and in Experiment 4 (β = -7.712, SE = 2.82, t = -2.74, $p = .007$) than in Experiment 1 (Figure 8b). Objective ostracism in Experiments 3 and 4 did not differ significantly (p = .76), although the difference was in the predicted direction (c.f., Figure 8a). Similarly, objective ostracism was lower also in the five very last periods in Experiment 3 (β = -0.95, SE = 0.47, t = -2.04, $p = .042$) and Experiment 4 (β = -1.08, SE = 0.47, t = -2.27, $p = .024$) than in Experiment 1. In these analyses we controlled for free-riding in the initial (i.e., 1 or 5) paired periods, which, as the model predicted, in itself increased objective ostracism (β = 7.93, SE = 1.81, t = 6.72, $p < .0001$). Thus, while free-riding in the initial paired periods predicted being ostracized later, our analysis accounts for this variance, and shows that for an average participant (i.e., at the sample mean of free-riding in the paired periods), imposing path dependence reduces objective ostracism. This reduction was brought about by a stronger tendency to form lasting, mutual dyads ("friendships"), rather than an increase in network interconnectedness between all group members (see Supplementary Note 1 and Supplementary Figure 7 for additional analyses).

Imposing path dependence also reduced the inequality in objective ostracism: both Experiments 3 (~.26) and 4 (~.25) had lower Gini coefficients (adjusted for free-riding in the paired periods) than Experiment 1 (bootstrapped 95 % CI for difference between

Experiment 1 and Experiment 3: [-0.11, -0.006], between Experiment 1 and Experiment 4 [-0.11, -0.02]. See Supplementary Figure 7 for Lorenz curves). Finally, we found that imposing path dependence led to a reduction also in the subjective experience of ostracism. Subjective ostracism was reduced in both Experiments 3 (β = -0.66, SE = 0.17, t = -3.78, $p < .001$) and 4 (β = -0.7, SE = 0.17, t = -4.0, $p < .0001$) relative to Experiment 1 (Figure 8c). Together, these results provide strong evidence for the RL basis of incidental ostracism, and demonstrate that path dependence, which in itself is a signature of incidental ostracism, can be leveraged as an intervention to reduce both objective and subjective ostracism.

Discussion

Our results demonstrate that a substantial proportions of ostracism can be incidental and emerge from simple reinforcement learning mechanisms. Thus, rather than being restricted to the traditionally assumed instrumental response to the free-riding of others, ostracism occurred in our experiments to a large degree in the absence of free-riding. The emergence account of ostracism offers a distinct view on how the basic components of social interactions - selecting partners and establishing relationships - can lead to a substantial degree of ostracism, both expressed in objective social network structure and subjective experience. Furthermore, it provides clues of how ostracism can be reduced. Note that our findings do not disprove previous theoretical and empirical demonstrations of the power of ostracism to prevent free-riding, but suggest that commonly observed ostracism might be less instrumental than previously thought.

Our use of agent-based computational modeling with empirically estimated parameters allowed us to demonstrate that basic RL mechanisms, without motivation for instrumental ostracism, are sufficient for explaining the results. In contrast, formalization of the instrumental account provided a worse fit to the experimental data, and failed to account for

path dependent ostracism. According to our model, path dependent ostracism originated from a positive feedback mechanism captured by RL, as consecutive rewarding interactions with a given partner cumulatively increase the probability of selecting the same partner again in the future, at the expense of selecting alternative partners. Naturally, this implies that path dependence can occur also in many non-social decision making situations where the environment provides structurally similar positive or negative feedback[38]. Our results further expand the scope of simple RL mechanisms for explaining seemingly complex social behaviors, such as economic game behavior[46], social network structure[37], behavioral traditions[47], and social cognition[48].

Although the agent-based model shows that instrumental ostracism is not necessary for explaining the experimental results, we cannot exclude that some participants did have instrumentally punitive motivations for their partner choices. However, our analysis puts a clear upper bound on instrumental ostracism in our experiments, by showing that free-riding is a weak predictor of ostracism when path dependence is accounted for (Figure 4) and that incidental ostracism emerges also in a social setting where free-riding is not an option (Experiment 2). An intriguing possibility for real-world ostracism is the interaction between incidental and instrumental ostracism, so that incidental outcasts become stigmatized, for example if observers view ostracism as a signal of deviance ("no smoke without fire"), which could result in subsequent instrumental ostracism of the same individuals.

Our findings have policy implications, as they provide an alternative explanation for the prevalence of ostracism, and make new suggestions for preventive interventions. Specifically, Experiments 3 and 4 demonstrated that incidental ostracism can be reduced by utilizing path dependence to enforce dyadic interaction between group members. The reduction of ostracism was brought about by an enhanced tendency for individuals to form stable mutual dyads ("friendships"), rather than an increased interconnectivity between all group-members.

Such mechanistically informed strategies represent important complements to previously suggested methods for preventing ostracism[3,4], for example in school classes.

Several limitations should be noted. Our experimental model, which was stylized and based on economic games with monetary payoffs, does not directly speak to forms of instrumental ostracism targeted at individuals who are not free-riders in a standard sense. For example, previous research has shown that people ostracize slow individuals in a virtual ball tossing game (Cyberball[49])[25,50]. Such individuals are not free-riders per se but burdensome, in that they cause frustration and reduce the per unit time payoff to the focal individual (who typically knew the total numbers of rounds in the experiment[25]). However, our emergence account may well generalize to such situations, as interacting with burdensome individuals likely results in lower than average payoffs, which in turn promotes avoidance. In any case, evaluating the emergence models in other experimental situations, involving non-monetary payoffs, represents an important future direction. Other forms of punitive ostracism, such as the silent treatment between spouses[28], fall outside our emergence account. Moreover, as our experiments were based on dyadic interactions, where the participants had no knowledge about the payoffs of group members they did not interact with, our findings do not speak to the role of welfare or group-level payoff concerns for the emergence of incidental ostracism. This intriguing question could for example be addressed by combining dyadic interaction with group interactions (e.g., in a Public goods game). Finally, given that our experimental model by design is highly simplified, it is likely that additional factors contribute to ostracism in real-world social situations (e.g., prior information about other individuals, physical appearance[51]). We view our experimental and computational model as a baseline to which such additional factors can be added.

In summary, our experiments and computational modeling provide new insights into the mechanistic causes of ostracism in groups of interacting people. Understanding

ostracism as an emergent phenomenon has important ramifications for the analysis of its social function, and for designing interventions to combat ostracism.

Methods

Participants. A total of 456 (259 female, mean age ~ 24) individuals participated in the experiments, and typically (median) earned 37 CHF (~ 36 US $) during the experiment. Each experimental session included 2-3 groups of six participants. Experiment 1 included 186 participants, and the three other experiments 90 participants each. The study was approved by the Human Subjects Committee of the Faculty of Economics, Business Administration, and Information Technology at the University of Zurich. Because the study was non-invasive and did not feature deception, signed consent was not required according to the approval. No individuals who had participated in any experiments involving deception (at the University of Zürich) were invited for participation.

Experimental task. The experiments were conducted at the Laboratory for Experimental and Behavioral Economics at the Department of Economics, University of Zürich. Participants were seated in separate "cubicles" in a larger room. None of the experiments involved any deception. Before the experiment, the experimenter provided a brief verbal introduction to the participants, stressing that the experiment was conducted in real-time. The participants played an iterated Prisoner's Dilemma (Experiments 1 & 3-4) or coordination (Experiment 2) game with partner choice for 35 periods in groups of six. The payoffs for the Prisoner's Dilemmas were C/C = 2, C/D = 0, D/C = 3, D/D = 1, where C = Cooperate and D = Defect. The payoffs for the coordination game was A/A = 3, A/B = 0, B/B = 3, B/A = 0, where A = Action A and B = Action B.

Each experimental period began with a partner choice stage (Figure 1) where each player was indicated with a number (e.g., "Player 1") in a matrix-like representation (see Supplementary Figure 1 for a period time-line and additional information). After partner choices, the participants were prompted to select an action (as initiator) for the interaction with the selected partner. The actions were referred to as action "A" and action "B" (e.g., rather than "Cooperate" and "Defect") to avoid pronounced framing effects. The participants were then sequentially notified about which other players had selected them as partner, and asked to indicate their action in this interaction (as responder). Next, the payoffs for the initiator and the responder were displayed on their respective screens for each selection (thus, the action of the interaction partners was not explicitly stated and participants inferred it from the simple payoff matrix). If more than one player had selected a given player as partner, these interactions were realized in randomized order. Players who had not been selected as partners during a period waited until all interactions had been realized. Participants were not notified about the number of periods, to prevent end-game effects. The experiments began with written instructions, specifying the payoff matrix and the game rules (see Supplementary Methods for instructions), which was followed by a quiz to verify understanding. After the experiments, the participants filled in a computerized version of the need-threat scale[39], a standard measure of subjective ostracism. The scale is composed of four subscales, designed to measure the fundamental psychological needs of belonging, self-esteem, meaningful existence, and sense of control. We averaged and reversed the subscales as a compact index of subjective ostracism.

Experiments 3 & 4 in addition included instructions that participants would be paired with another player for one period (Experiment 3) or an undisclosed (to prevent "end" game effects) number of periods (Experiment 4) before they would be free to select partners for the Prisoner's Dilemma interaction. Instructions and procedures of these two experiments were kept as close to Experiment 1 as possible.

Experiment1 1 had two payoff conditions: the cumulative payoff condition (n = 78) and the randomized payoff (n = 108) condition. In the former, the payoff from each Prisoner's Dilemma interaction in each period was summed up. In the latter, the payoff from one interaction in each period was randomly selected. To balance the monetary profit between the conditions, the randomized payoff condition had a five times higher conversion rate from experimental currency units to CHF. The randomized payoff condition was intended to mimic natural variability in the reinforcement provided for one´s behavior by the social environment as a test of robustness. Incidental ostracism generalized across the two reinforcement environments (Supplementary Note 1). Experiments 2-4 were all based on the randomized payoff condition,

The payoff condition was crossed with a condition where the popularity (i.e., in-degree) of each player from the previous period was displayed during the partner choice stage (i.e., visible [n = 96] vs not visible [n = 90]). This condition had no discernable effects on the results, and is for this reason only reported in Supplementary Table 6 and in the Supplementary Note 2. The reported results are based on data collapsed across this condition. Experiments 2-4 involved no social information (i.e., the participants could not observe the popularity of the other players).

Statistical analysis. All analyses were conducted in R 3.31. Linear regression was performed with the *lm* function. Explained variance refers to adjusted R^2. See Supplementary Tables 3-6 for alternative analyses using Quasi-Poisson regression. Bootstrap confidence intervals are percentile intervals.

Data availability. The experimental data is available from the corresponding author.

Real-world social networks. See Supplementary Methods for details about the real-world social network data analysis.

Agent-based modeling and model comparison. See Supplementary Methods for detailed information about the agent-based models and model-comparison.

Code availability. The modeling code is available from the corresponding author.

References

1. Williams, K. D. Ostracism. *Annu. Rev. Psychol.* **58,** 425–452 (2007).

2. Wesselmann, E. D. & Williams, K. D. Social life and social death: Inclusion, ostracism, and rejection in groups. *Gr. Process. Intergr. Relations* **20,** 693–706 (2017).

3. Robinson, S. L. S., O'Reilly, J. & Wang, W. Invisible at Work An Integrated Model of Workplace Ostracism. *J. Manage.* **39,** 203–231 (2013).

4. Williams, K. D. & Nida, S. A. Ostracism and Public Policy. *Policy Insights from Behav. Brain Sci.* **1,** 38–45 (2014).

5. Blackhart, G. C., Eckel, L. A. & Tice, D. M. Salivary cortisol in response to acute social rejection and acceptance by peers. *Biol. Psychol.* **75,** 267–276 (2007).

6. Eisenberger, N. I., Lieberman, M. D. & Williams, K. D. Does rejection hurt? An FMRI study of social exclusion. *Science* **302,** 290–2 (2003).

7. Wesselmann, E. D., Nairne, J. S. & Williams, K. D. An evolutionary social psychological approach to studying the effects of ostracism. *J. Soc. Evol. Cult. Psychol.* **6,** 309–328 (2012).

8. Sasaki, T. & Uchida, S. The evolution of cooperation by social exclusion. *Proc. R. Soc. London B Biol. Sci.* **280,** (2012).

9. Kurzban, R. & Leary, M. R. Evolutionary origins of stigmatization: the functions of

social exclusion. *Psychol. Bull.* **127,** 187–208 (2001).

10. Gruter, M. & Masters, R. D. Ostracism as a social and biological phenomenon: An introduction. *Ethol. Sociobiol.* **7,** 149–158 (1986).

11. Nakamaru, M. *et al.* The Effect of Ostracism and Optional Participation on the Evolution of Cooperation in the Voluntary Public Goods Game. *PLoS One* **9,** e108423 (2014).

12. Johnson, T. The strategic logic of costly punishment necessitates natural field experiments, and at least one such experiment exists. *Behav. Brain Sci.* **35,** 31–2 (2012).

13. Bowles, S. & Gintis, H. The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theor. Popul. Biol.* **65,** 17–28 (2004).

14. Wesselmann, E. D., Wirth, J. H., Pryor, J. B., Reeder, G. D. & Williams, K. D. When Do We Ostracize? *Soc. Psychol. Personal. Sci.* **4,** 108–115 (2013).

15. Guala, F. Reciprocity: weak or strong? What punishment experiments do (and do not) demonstrate. *Behav. Brain Sci.* **35,** 1–15 (2012).

16. Boehm, C. *et al.* Egalitarian Behavior and Reverse Dominance Hierarchy [and Comments and Reply]. *Curr. Anthropol.* **34,** 227–254 (1993).

17. Hirshleifer, D. & Rasmusen, E. Cooperation in a repeated prisoners' dilemma with ostracism. *J. Econ. Behav. Organ.* **12,** 87–106 (1989).

18. Feinberg, M., Willer, R. & Schultz, M. Gossip and ostracism promote cooperation in groups. *Psychol. Sci.* **25,** 656–64 (2014).

19. Maier-Rigaud, F. P., Martinsson, P. & Staffiero, G. Ostracism and the provision of a

public good: experimental evidence. *J. Econ. Behav. Organ.* **73,** 387–395 (2010).

20. Cinyabuguma, M., Page, T. & Putterman, L. Cooperation under the threat of expulsion in a public goods experiment. *J. Public Econ.* **89,** 1421–1435 (2005).

21. Davis, B. J. & Johnson, D. B. Water Cooler Ostracism: Social Exclusion as a Punishment Mechanism. *East. Econ. J.* **41,** 126–151 (2015).

22. Hales, A. H., Kassner, M. P., Williams, K. D. & Graziano, W. G. Disagreeableness as a Cause and Consequence of Ostracism. *Pers. Soc. Psychol. Bull.* **42,** 782–97 (2016).

23. Kagel, J. & McGee, P. Personality and cooperation in finitely repeated prisoner's dilemma games. *Econ. Lett.* **124,** 274–277 (2014).

24. Wesselmann, E. D., Wirth, J. H., Pryor, J. B., Reeder, G. D. & Williams, K. D. The Role of Burden and Deviation in Ostracizing Others. *J. Soc. Psychol.* **155,** 483–496 (2015).

25. Wesselmann, E. D., Williams, K. D. & Wirth, J. H. Ostracizing Group Members Who Can (Or Cannot) Control Being Burdensome. *Hum. Ethol. Bull.* 82–103 (2014).

26. Wirth, J. H., Bernstein, M. J. & LeRoy, A. S. Atimia: A New Paradigm for Investigating How Individuals Feel When Ostracizing Others. *J. Soc. Psychol.* **155,** 497–514 (2015).

27. Williams, K. D. Ostracism: The Kiss of Social Death. *Soc. Personal. Psychol. Compass* **1,** 236–247 (2007).

28. Nezlek, J. B., Wesselmann, E. D., Wheeler, L. & Williams, K. D. Ostracism in everyday life. *Gr. Dyn. Theory, Res. Pract.* **16,** 91–104 (2012).

29. Yang, J. & Treadway, D. C. A Social Influence Interpretation of Workplace Ostracism

and Counterproductive Work Behavior. *J. Bus. Ethics* 1–13 (2016). doi:10.1007/s10551-015-2912-x

30. Nezlek, J. B., Wesselmann, E. D., Wheeler, L. & Williams, K. D. Ostracism in Everyday Life: The Effects of Ostracism on Those Who Ostracize. *J. Soc. Psychol.* **155,** 432–451 (2015).

31. Holland, J. *Emergence: From chaos to order*. (Addison-Wesley, 1998).

32. Sommer, K. L., Williams, K. D., Ciarocco, N. J. & Baumeister, R. F. When Silence Speaks Louder Than Words: Explorations Into the Intrapsychic and Interpersonal Consequences of Social Ostracism. *Basic Appl. Soc. Psych.* **23,** 225–243 (2001).

33. Rand, D. G., Arbesman, S. & Christakis, N. A. Dynamic social networks promote cooperation in experiments with humans. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 19193–8 (2011).

34. Shirado, H., Fu, F., Fowler, J. H. & Christakis, N. A. Quality versus quantity of social ties in experimental cooperative networks. *Nat. Commun.* **4,** 2814 (2013).

35. Wang, J., Suri, S. & Watts, D. J. Cooperation and assortativity with dynamic partner updating. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 14363–8 (2012).

36. Barclay, P. Biological markets and the effects of partner choice on cooperation and friendship. *Curr. Opin. Psychol.* **7,** 33–38 (2016).

37. Skyrms, B. & Pemantle, R. A dynamic model of social network formation. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 9340–6 (2000).

38. Page, S. E. Path Dependence. *Quart. J. Polit. Sci.* **1,** 87–115 (2006).

39. Zadro, L., Williams, K. D. & Richardson, R. How low can you go? Ostracism by a

computer is sufficient to lower self-reported levels of belonging, control, self-esteem, and meaningful existence. *J. Exp. Soc. Psychol.* **40,** 560–567 (2004).

40. Grimm, V. *et al.* Pattern-Oriented Modeling of Agent-Based Complex Systems: Lessons from Ecology. *Science (80-.).* **310,** (2005).

41. Smith, E. R. & Conrey, F. R. Agent-based modeling: a new approach for theory building in social psychology. *Pers. Soc. Psychol. Rev.* **11,** 87–104 (2007).

42. Rescorla, R. A. & Wagner, A. in *Classical conditioning II: Current research and Theory* (eds. Black, A. H. & Prokasy, W. H.) 64–99 (Appleton-Century-Crofts, 1972).

43. Finite automata play the repeated prisoner's dilemma. *J. Econ. Theory* **39,** 83–96 (1986).

44. Palminteri, S., Wyart, V. & Koechlin, E. The Importance of Falsification in Computational Cognitive Modeling. *Trends Cogn. Sci.* **21,** 425–433 (2017).

45. Steingroever, H., Wetzels, R. & Wagenmakers, E.-J. Absolute performance of reinforcement-learning models for the Iowa Gambling Task. *Decision* **1,** 161–183 (2014).

46. Erev, I. & Roth, A. E. Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria. *Am. Econ. Rev.* **88,** 848–81 (1998).

47. Lindström, B. & Olsson, A. Mechanisms of social avoidance learning can explain the emergence of adaptive and arbitrary behavioral traditions in humans. *J. Exp. Psychol. Gen.* **144,** 688–703 (2015).

48. Zaki, J., Kallman, S., Wimmer, G. E., Ochsner, K. & Shohamy, D. Social Cognition as Reinforcement Learning: Feedback Modulates Emotion Inference. *J. Cogn. Neurosci.*

28, 1270–1282 (2016).

49. Hartgerink, C. H. J., Beest, I. Van, Wicherts, J. M. & Williams, K. D. The Ordinal Effects of Ostracism : A Meta- Analysis of 120 Cyberball Studies. *PLoS One* 1–24 (2015). doi:10.1371/journal.pone.0127002

50. Wesselmann, E. D., Wirth, J. H., Pryor, J. B., Reeder, G. D. & Williams, K. D. The Role of Burden and Deviation in Ostracizing Others. *J. Soc. Psychol.* **155,** 483–496 (2015).

51. Janssen, I., Craig, W. M., Boyce, W. F. & Pickett, W. Associations between overweight and obesity with bullying behaviors in school-aged children. *Pediatrics* **113,** 1187–94 (2004).

Correspondence should be addressed to Björn Lindström, bjorn.lindstrom@econ.uzh.ch

Author contributions

B.L. and P.N.T. conceived the study and designed the experiments. B.L. collected the data. B.L. developed the models. B.L. analyzed the data and implemented the models. B.L. and P.N.T. wrote the paper.

Acknowledgments

We thank Charles Efferson for valuable suggestions regarding the design and implementation of experiment 1, and Andreas Olsson, Philip Pärnamets and Ida Selbing for helpful comments on an earlier version of the manuscript. This research was supported by Swiss NSF grants PP00P1_150739, 00014_165884, and 100019_176016 to Philippe N. Tobler. Björn Lindström

was supported by Forte (COFAS2: 2014-2785 FOIP). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests

The authors declare no competing interests.

Figure legends

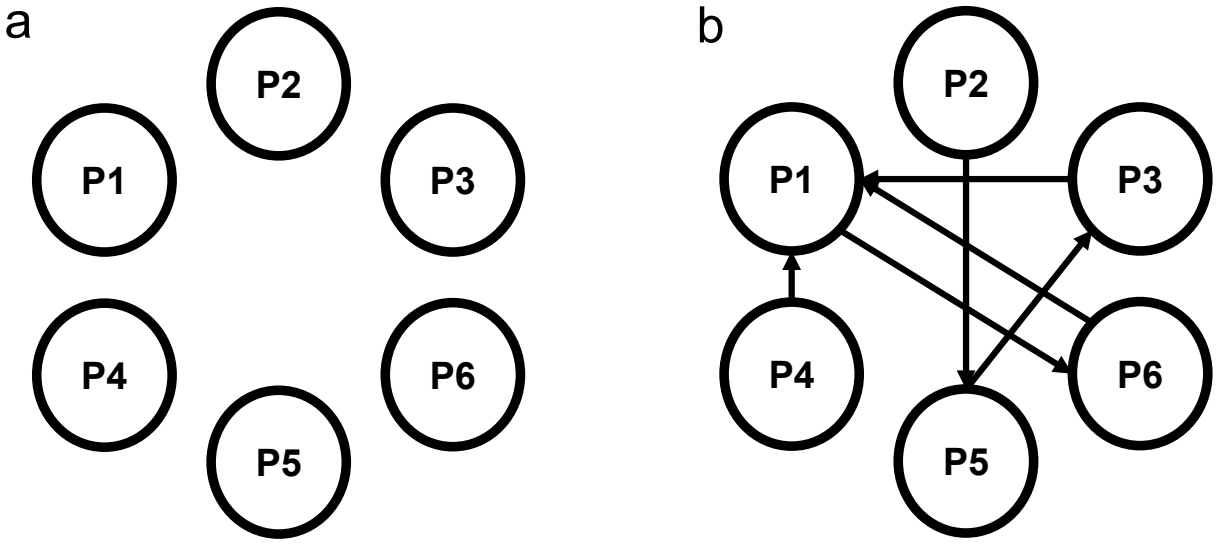

Figure 1. Experimental design. Participants played a repeated Prisoner's Dilemma (Experiments 1 & 3-4) or coordination game (Experiment 2) with free partner choice in groups of six players. (**a**) At the outset of a period, each participant chose a partner. (**b**) In each period, each player could take part in up to six independent games: one with the partner s/he selected (as the initiator) and up to five with the players who selected him/her (as the responder). For example, player 1 (P1) selected player 6 (P6) as initiator and was selected as partner by three players, resulting in four games for P1 (one as initiator and three as responder), while player 4 (P4) was not selected as partner by any other player, resulting in one game for P4 (as initiator) in that period. The players made separate game choices for each interaction, allowing each action to be conditioned on previous interactions (e.g., P1 could decide to cooperate with P6 but not with P4). The players only knew about the choices of players they interacted with themselves (e.g., P1 knew that P3 cooperated with P1, but not that P3 also interacted with P5, or the action choices of P3 and P5 in their interaction). The experiment consisted of 35 periods in total. See Methods for

information about the game payoffs, and Supplementary Figure 1 for a detailed overview of an experimental period from the point of the participant.

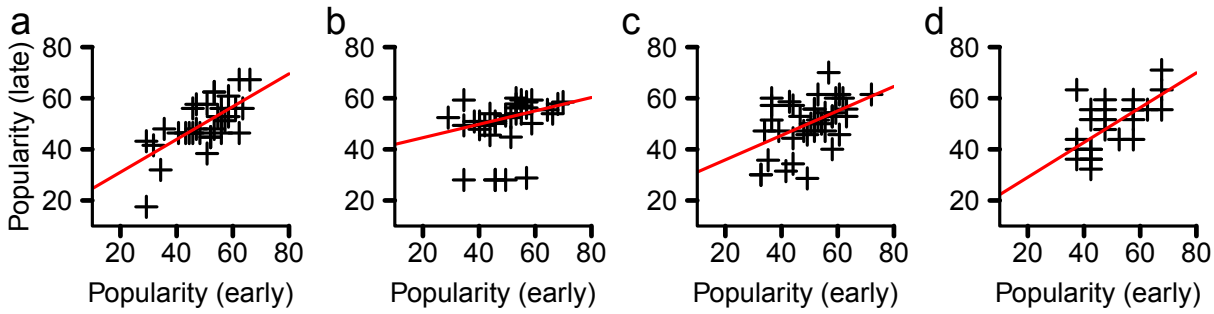

Figure 2. Prediction of late popularity by early popularity in four different real-world social networks. (**a**) University freshmen class, n = 32, (**b**) University freshman class, n = 34, (**c**) University freshmen class, n = 38, (**d**) Secondary school class, n = 26. Each cross denotes an individual. The red lines indicate the best fitting robust regression slopes. The popularity scores are T-scored (standardized with M = 50 and SD = 10) for comparability.

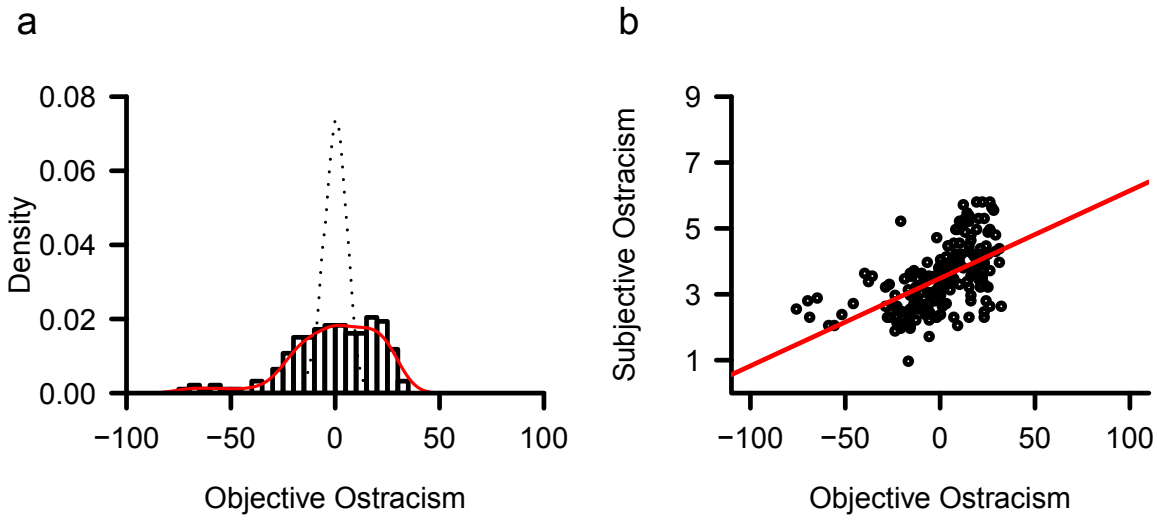

Figure 3. Objective and subjective ostracism in Experiment 1. (**a**) **Objective ostracism is prevalent**. Positive values denote participants who were selected as partner on average less than once per period, and negative values denote participants who were selected as partner on average more than once per period. The dotted line represents the theoretical reference distribution based on completely random choices (i.e., absence of ostracism). The red line represents the empirical density. The figure includes all participants (n = 186) in Experiment 1. (**b**) **Objective**

ostracism predicts subjective ostracism. The red line shows the unadjusted slope of objective ostracism. The figure includes all participants (n = 186) in Experiment 1. Each point represents one individual.

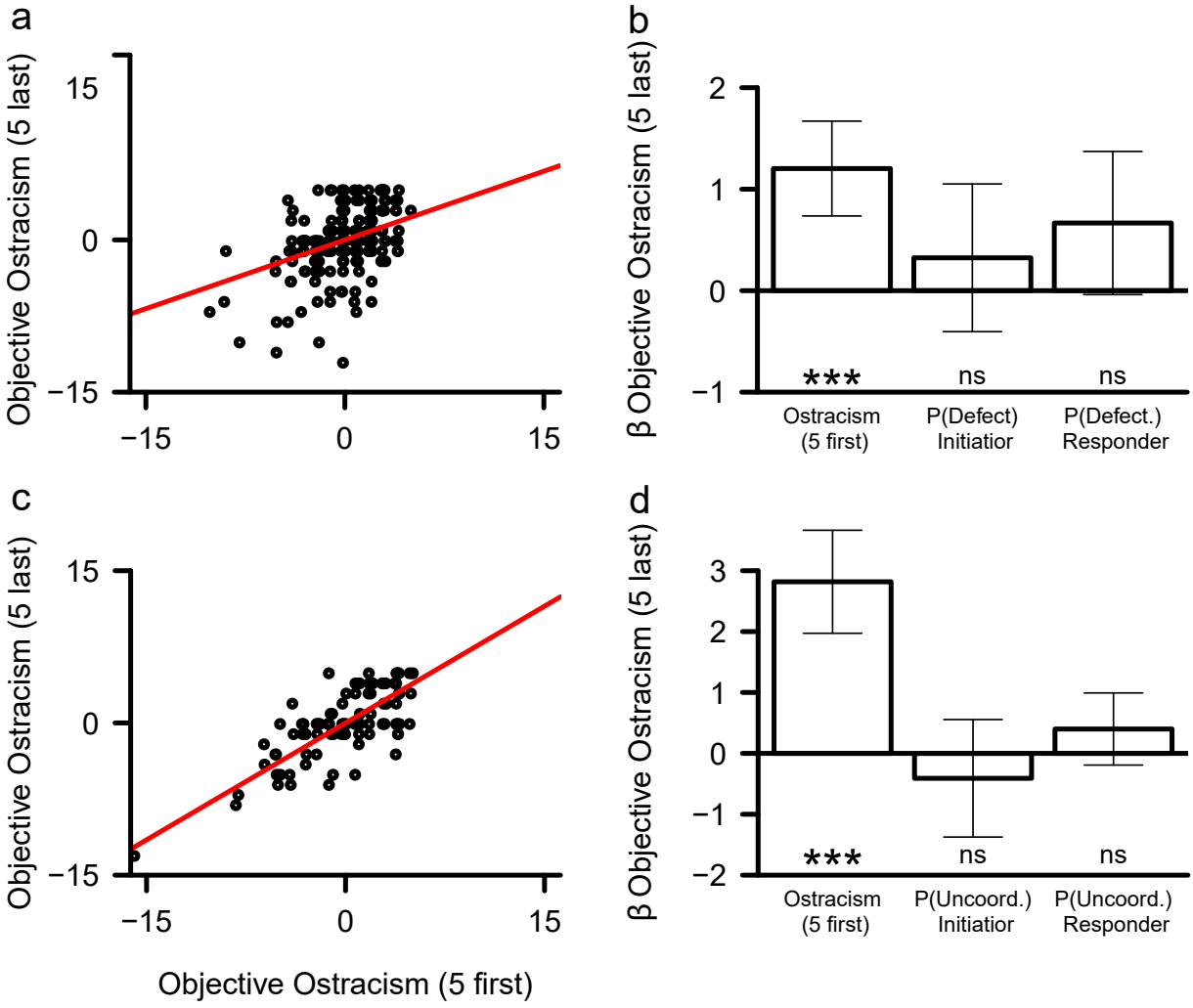

Figure 4. Incidental, path-dependent ostracism in the Prisoner's Dilemma game of Experiment 1 (n = 186) (a-b). (**a**) **Positive relationship between early (five first) and late (five last) periods indicative of ostracism being path dependent**. The solid line shows the slope, adjusted for the two free-riding regressors (% Defect actions as initiator and Responder). The points are slightly jittered along the x-axis for higher visibility. The figure includes all data points (n = 186). (**b**) **Early ostracism predicts late ostracism better than free-riding.** Contribution of each regressor (standardized) to the prediction of objective ostracism during the last five periods (model total R^2 = .29**)** (see text for regression model details). **Incidental, path dependent ostracism in the pure coordination game of Experiment 2 (n = 90) (c-d).** (**c**) **Path dependence between the first and last five periods**. The solid line shows the slope, adjusted for the two coordination-failure regressors. The points are slightly jittered

along the x-axis for higher visibility. (**d**) **Early ostracism predicts late ostracism better than coordination failure.** Contribution of each regressor (standardized) to prediction of objective ostracism during the last five periods (model total R^2 = .54) (see text for regression model details). P(Uncoord.) = % of coordination failures. Error bars are 95% CI. *** = p < .0001, ns. = not significant.

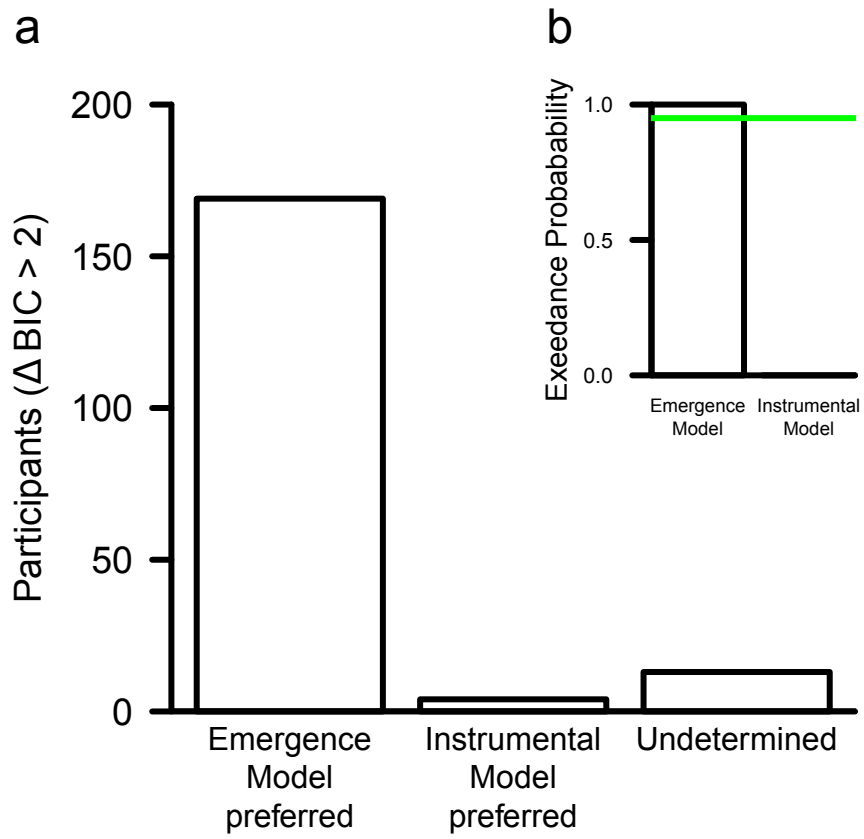

Figure 5. Comparison between emergence and instrumental models in Experiment 1. (**a**) **Individual participant level**. Choices of a large majority of participants were better explained by the emergence model than the instrumental model. A Bayesian information criterion (BIC) difference > 2 indicates positive evidence for one model over the other. (**b**) **Exceedance probability**. Random effects model comparison showing the posterior probability that the emergence model explains the data better. The green line displays P = 0.95.

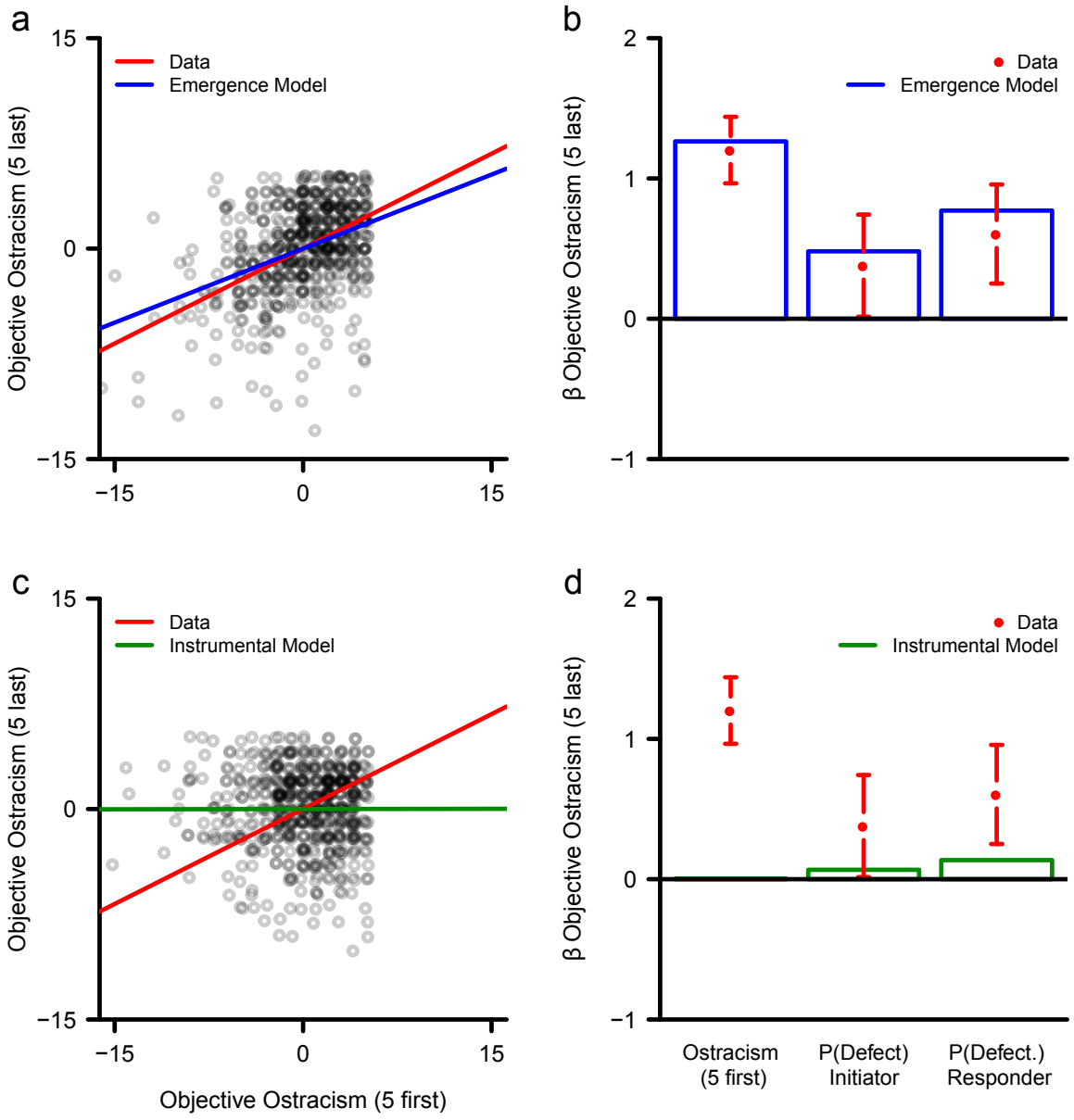

Figure 6. Model validation based on generative performance. (a & c) Simulated path dependence. For the simulations, we used the estimated parameter values from Experiment 1 and tested if the candidate models could generate path dependent ostracism. Dots are simulated data, lines show the slope of path dependence from the data (red), the emergence model (blue), and the instrumental model (green). **(b & d) Predictors of ostracism in simulated data (regression estimates)** (blue = emergence model, green = instrumental model) overlaid with parameter estimates from the experimental data (red, see Figure 4). Error bars are standard errors of the regression coefficients from the experimental data (c.f. Figure 4).

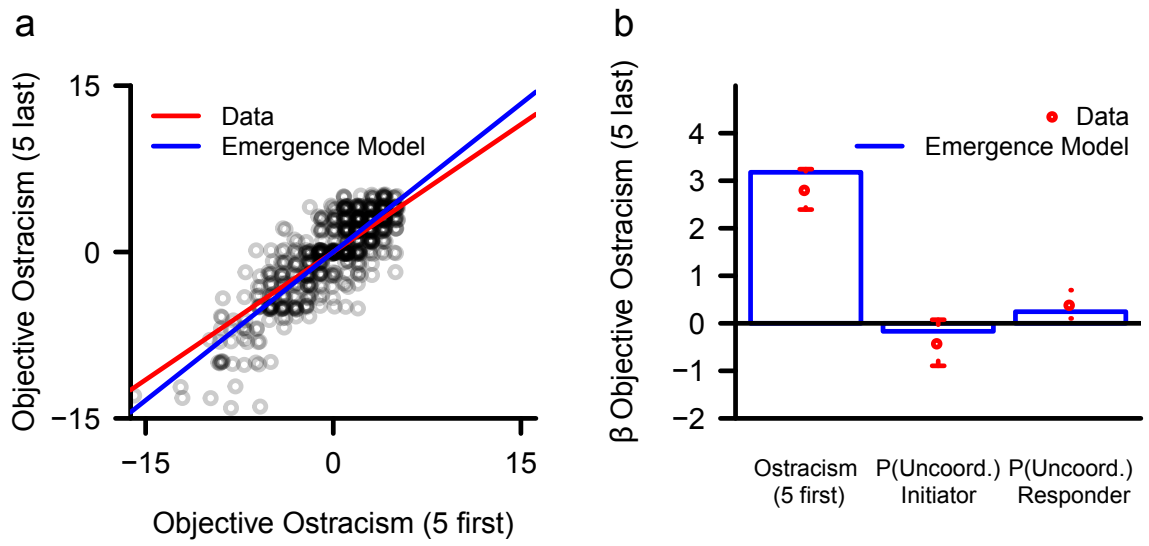

Figure 7. Model generalizability. (a) Path dependence in Experiment 2 (coordination game) simulated by emergence model derived from Experiment 1 (Prisoner's Dilemma). Dots are out-of-sample model predictions, based on parameter estimates from Experiment 1, lines show the slope of path dependence estimated from the experimental data (red) and the emergence model predictions (blue). (**b**) **Predictors of ostracism (regression estimates)** from emergence model predictions (blue) overlaid with parameter estimates from experimental data (red, see Figure 5). Error bars are standard errors of the regression coefficient from the experimental data.

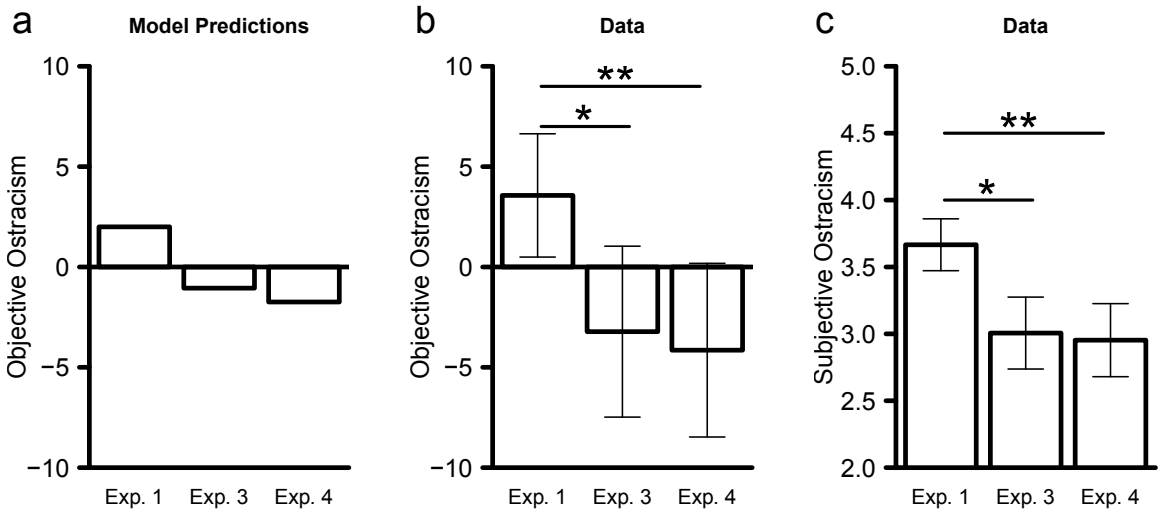

Figure 8. Reduction in objective and subjective ostracism by causal manipulation of path dependence in Experiments 3 and 4. (**a**) **Model predictions.** According to the emergence model, imposing path dependence through pairing each participant with another participant for the first period (Experiment 3) or the first five periods (Experiment 4) should reduce objective ostracism. The predictions are adjusted for the influence of free-riding during the paired periods. (**b**) **Confirmation of model predictions for objective ostracism**. In keeping with the emergence model, objective ostracism was lower in Experiments 3 and 4 compared to Experiment 1. The plot shows regression coefficients, adjusted for free-riding during the paired periods. Error bars are 95% CI. (**c**) **Imposing path dependence reduces subjective ostracism.** The plot shows regression coefficients, adjusted for free-riding during the paired periods. Error bars are 95% CI. * = $p < .05$, ** = $p < .01$.