4

The Double-Edged Sword of Cooperation

The Vulnerability of Trusting and Trustworthy Behaviour

BURAK SONMEZ

In situations where personal interests are particularly at odds with collective interests, trust is central to understanding cooperation. Since societies' dependence on cooperation in providing public goods grows incrementally in physical and virtual environments (eg, online petitions, open-access digital archives and online crowdsourcing), we are (unsurprisingly) witnessing the need for trust as a prerequisite for cooperation. In practice, individuals or entities are frequently protected by laws (*legal central authorities*) or *private ordering* to deal with free-riding in collective actions. Some scholars, for instance, show that individuals are willing to incur a small cost to punish free riders (Fehr and Gächter 2000; Fehr and Gächter 2002). However, how is it possible to encourage people to contribute to the provision of public goods in the absence of enforcement? It is particularly pivotal to identify the causal mechanism of cooperative behaviour through *learning, reliability, strategies or goodwill* when sanctions are unconvincing and costly to ensure public goods. In this vein, even though there is no clear-cut understanding of how trust and trustworthiness drive collective actions (for further discussions, see Bacharach et al 2007; Coleman 1994), scholars from different disciplines have sought to conceptualise and explain the fundamentals of how trust and cooperation go hand in hand (Buskens and Raub 2002; Gächter et al 2004; Gambetta 1988; J Jordan et al 2015; J Jordan et al 2016; Ledyard 1995; Lo Iacono and Sonmez 2021; Rigdon et al 2007; Rothstein 2000; Uslaner 2002).

Much of the social sciences literature on trust suggests that the concept of trust has been formulated in multiple ways. For instance, there are two broadly defined views of trust: affective- and cognitive-based. Affective-based trust is considered interpersonal and emphatic concern towards others, while cognitive-based trust is conceptualised as a rational assessment of beliefs about reliability, strategies, credibility, and competence (McAllister 1995; Uslane 2002). Following this distinction, affective trust is also considered the willingness of an individual (the trustor) to be vulnerable to the actions of other(s) (the trustee) because the emotional bond between the trustee and trustor assures that the trustor's concern for the trustee is reciprocated. On the other hand, cognitive trust primarily relates to the willingness to be vulnerable to the actions of the trustee, determined by direct or indirect interactions with the trustee, such as learning about their reputation (McAllister, 1995). Scholars also define these vulnerabilities as some sort of expectations that the trustee will engage in the trusting behaviour even in the absence of an authority monitoring the actions (Colquitt et al 2007; Dirks and Ferrin 2002; Ho and Weigelt 2005; Montgomery et al 2008; Rousseau et al 1998).

Following such definitions in the literature, it is argued that individuals hold differential expectations of trust through their attitudes and information on others' reputation. People often hold and update these expectations unconsciously through their socialisation and interpersonal interactions. Relatedly, the greater people's expectations that their trusting behaviour will be honoured by trustees, the more vulnerable they become against the betrayal of trustees. However, this relational understanding between trust and vulnerability is often conceptualised through the lens of the trustor but not the trustee in the social sciences literature. In addition, this type of vulnerability predominantly focuses on expectations rather than actual structural vulnerability. For instance, status characteristics theory (STC) suggests that status differences between individuals (eg, refugees as low-status and citizens as high-status) in collective actions may create certain expectations where higher-status group members are expected to be more influential (Ridgeway 1991). Following the STC theory, I argue that the vulnerability of low-status groups in trusting relationships is structurally embedded, thereby leading to relatively unmalleable expectations on their trusting or trustworthy behaviour. Given the impact of roles (trustor vs trustee) and structural differences in the relationship between vulnerability and trust, there is bourgeoning literature on to what extent the current understanding of trust and trustworthiness through vulnerability is limited, especially when individuals have greater action spaces and different payoff structures in alternative Trust games than in the typical Trust games[1] (Cox et al 2016; Keser and Özgümüs 2018; Engler et al 2018). That is, scholars found that the trustor becomes more discouraged to send any amount if their actions become more vulnerable to the second mover's betrayal (Keser & Özgümüs, 2018). On the other hand, if the trustor decides to accept a set of opportunities with a high risk of being betrayed by the trustee, the trustee might see this intention-based social preference as a strong, trusting signal (Engler et al 2018).

[1] The most popular forms of Trust games are known as either investment or trust games (Berg et al 1995), where two players–a trustor and a trustee–start with identical initial endowments (ε). As a first step, the trustor chooses to transfer the amount $\theta \in [0, \varepsilon]$ to the trustee, and the amount sent is multiplied by γ before the trustee receives–conventionally, $\gamma > 1$. In the final stage, the trustee decides on whether to return the amount $\lambda \in [0, \theta\gamma]$ back to the trustor. Scholars, in general, consider the trustor as trusting if they sends more than the 'minimum' amount, and defines the trustee as trustworthy if they returns more than the 'minimum' amount or more than the amount sent to them (Chaudhuri and Gangadharant, 2007; Ben-Ner and Putterman, 2009).

In the following, I will consider different causal explanations to understand people's willingness to cooperate in collective actions where trust and trustworthiness are subject to various types of vulnerability. In addition, I will revisit the paradigmatic designs in understanding cooperation to discuss to what extent reputational information constructs expectations in trusting relationships and how vulnerability as both relational and structural concept is ignored in these designs. In doing so, the chapter will also shed some light on the challenges in establishing social cooperation between vulnerable groups, such as asylum seekers or refugees and host populations, in the absence of legal regulations.

I. Trust and Cooperation

The questions of whom I should trust, to what extent I can trust, and why I should trust become salient in revealing how much trust is required to sustain or engage in cooperation, especially when central authority and monitoring are deficient or non-existent. Scholars have shown that enforcement strategies for cooperation may actually backfire since the existence of institutional enforcement can motivate people to think that others cooperate to avoid punishment rather than their own goodwill or moral responsibility (Bowles 2008; Fehr and Rockenbach 2003; Mulder et al 2006). Therefore, once the enforcement or contractual obligation is no longer available, individuals may be more likely to free ride on the cooperation of others.

On the one hand, tackling global challenges such as humanitarian crises, climate change and health crises urges countries to cooperate to alleviate the dire impacts of these crises. On the other, the standard game-theoretic assumption is that cooperation across populations in contributing to a common resource (the public good) is implausible when the dominant strategy is to free ride on the contribution of others (Kollock 1998). Studying how to solve social dilemmas, Kollock categorised them into three broad groups: *motivational, strategic*, and *structural*. Motivational solutions are related to other-regarding preferences by which individuals are concerned about their own payoffs and others' gains (for reviews, see *social value orientation*, Van Lange 1999). However, I am more interested in focusing on strategic and structural solutions where social conditions (eg, social learning) and structural mechanisms (eg, reputation systems) highlight the conceptual salience of trust in cooperation. Individuals with a strong incentive to protect their reputation, for instance, are considered trustworthy in cooperative interactions (Diekmann et al 2014; Kuwabara 2015; Milinski 2016; Przepiorka et al 2017; Przepiorka and Berger 2017; Yamagishi et al 2009). Given the development of trust between particular groups or individuals, such as neighbours, co-workers and online community members over repeated interactions, scholars also examine how individuals extrapolate from their experiences in dyadic interactions to establish certain trust-related expectations in similar or different cooperative contexts (Glanville 2004; Glanville and Paxton 2007; Lo Iacono and Sonmez 2021). Another explanation to understanding why people avoid free-riding among group members, irrespective of enforcement mechanisms, is status hierarchies. Individuals, who contribute more to the group, may be regarded as *higher status people* since the contributions to public goods determine one's ranking in relation to the social hierarchy (Flynn et al 2006; Milinski et al 2002). High contributors are considered to be concerned with group welfare (Ridgeway 1978; 1982). From this point of view, they are highly expected to comply with the moral obligation of being honourable in cooperating with low-status individuals (Homans 1958). Taken together, they seem to signal more trustworthiness while exerting greater influence over other group members. In that vein, individuals are more likely to cooperate with high contributors, even in the situation where high contributors are more vulnerable to exploitation (Willer 2009). However, none of these studies has examined to what extent vulnerability embedded in social interactions and structural mechanisms influences trusting and trustworthy behaviour and, eventually, cooperation. The next section delves into understanding the juxtaposition of trust and vulnerability.

II. Trust and Vulnerability

In cooperating with others, individuals are exposed to specific uncertainties and vulnerabilities because of their trusting and trustworthy actions. Accordingly, trust is considered a *relational risk* where individuals are willing to engage in a social exchange by accepting the vulnerability to the actions of others based on the expectations about the others (Rousseau et al 1998). In our everyday life, we often think of vulnerability with negative connotations as an unfavourable condition that needs to be alleviated immediately. However, at the same time, we often voluntarily put ourselves into such situations where we fall in love or gossip about someone or rely on other drivers at the traffic light, thereby becoming vulnerable to others' failures to comply with our trusting or trustworthy actions. Individuals are willing to accept vulnerability to a certain extent in social exchanges to gain some emotional or material payoffs. Another strand of the literature focuses on the fact that the trustor has to accept their vulnerability in a *leap of faith* based on the goodwill of others rather than calculated expectations of trustworthiness. Notably, early sociological studies (see Georg Simmel 1950) distinguish one type of trust, in which the element of faith towards a stranger is beyond knowledge or particular reasons and combined with a sort of transcendent conception (Möllering 2006).

Rather, I reflect on the conventional notion that a trustor expects a trustee to honour the trusting act because the expected outcome with this social exchange outweighs the uncertainty of their vulnerability to the trustee. Following this line of thought, I first suggest that this conception of trust has been predominantly considered relational between the trustor and the trustee in the empirical studies, and often assumes that a trustor is vulnerable to a trustee's failure to honour, yet the fact that the trustor has the *structural advantage* to reserve the right not to initiate trusting behaviour is underexamined. Elaborating on what I define as the structural advantage, I stress that structural vulnerability entails a contextual dimension in trust-building processes, in which expectations of vulnerability to one's action are hardly independent of variances in status hierarchies. Those status differences, in general, drive interpersonal perception and behaviour in social contexts (Ridgeway and Berger 1986).

In a typical Investment or Trust Game studied by (Berg et al 1995; Kreps, 1990), a player has no control over which role (*sender or receiver*) they will be assigned. Player$_i$ is assigned to the sender role, while player$_j$ is assigned to the receiver role with probability p and the reverse allocation with probability $1 - p$. After the roles are randomly allocated, the trustor (sender) moves first to decide whether or not to trust the trustee (receiver) with a given endowment, x amount, before the trustee decides whether or not to reciprocate. Summarising the game accordingly, when the game is played only once (one round), the action set for the trustor is {Trust, Not trust}, and the action set for the trustee is {Reciprocate, Not reciprocate}. If the game is played for one round–not conditional on player's past interactions–the role-contingent strategies to cooperate could be interpreted as unconditional affective trusting behaviour that suggests a moral obligation to trust and expect reciprocity rather than betrayal, even though traditional game theory predicts that trust and reciprocity are not sustainable in finite durations (Olson 2012). On the other hand, in the repeated interactions, it is assumed that people make their best decision on whether to trust or not, relying on their most recent belief from the interactions (eg, see individual learning theories Buskens and Raub 2002; Skyrms 1990). Thus, in the repeated interactions, the trustee could have an opportunity in their action set to change the trustor's prior probability on the belief that the trustee will not betray.

However, one caveat in these games is that the trustee cannot take any action to challenge the trustor, who does not give any opportunity to entail a possibility of betrayal. In contrast to the limited action space in these games, the most recent literature on experimental economics shows that the structure of the game itself can modulate the expectations of the actors on what norms apply to the particular situation (Keser and Özgümüs 2018; List 2007). If the trusted actor also has an opportunity to take money away from the trustor, as well as the amount received, the trustor may expect less trustworthiness because of the increased vulnerability to the actions of the trustees, thereby trusting less. On the other hand, Engler et al (2017) show that when the trustor chooses an action space where the trustee can cause a potential loss for her, the trustee interprets this choice as a signal of enhanced vulnerability and performs in a more trustworthy manner in turn. This growing literature taken together implies that trusting and trustworthy behaviours may foster cooperation if the degree of vulnerability is determined by the trustor rather than imposed by a third party. Given this structural dynamic between vulnerability and trusting behaviour, it is important to examine how one's place in status hierarchies may modulate the expectations of being vulnerable to others' actions. That is, people in higher status positions expect more honouring behaviour than those in lower status positions (Berger et al 1995). These status characteristics, eg, become concrete when disadvantaged group members believe that most people perceive their group to be less competent and less trustworthy (Ridgeway 1997). Relatedly, one may further question whether informal status hierarchies naturally occurring through repeated interactions among individuals and dictated hierarchies would lead to different expectations of vulnerability in trusting relationships.

Reflecting on the status hierarchies in interpersonal interactions, I contend that the conventional operationalisation of trust through Trust games does not account for the variation in trustee's or trustor's prior expectations about their future interactions, which stem from their status position. Therefore, researchers should also disentangle to which extent people's expectations of vulnerability to one's action in trusting relationships derive from status characteristics, as well as their roles and dyadic interactions. In building trusting relationships in real-world social dilemmas, status hierarchies of both trustees and trustors are often signalled or visible to each other. For instance, Falk and Zehnder (2013), using a Trust game in a field experiment, show that individuals tend to invest more money (placing more trust) into the residents of the districts characterised by higher socioeconomic status rather than those living in the districts affiliated with lower socioeconomic status. Similarly, Trifiletti and Capozza (2011) found that in a Trust game, participants from Southern Italy characterised as low status in relation to participants from Northern Italy trusted the high-status participants from Northern Italy more than their own group. Following this reflection, people's expectations of vulnerability to trustee's or trustor's action are not only influenced by one's reputation based on personal interactions but also by their social status. In this regard, the following section discusses why a traditional causal pathway of reputational formation in interpersonal trust to sustain cooperation may mislead researchers to resolve social dilemmas.

III. Reconsidering the Paradigms in Cooperation

Traditional causal pathway to study cooperation through trust between host populations and identified vulnerable aliens, such as refugees, may be confounded. For example, in the conventional causal paths, a stranger is expected to cooperate or help if the other actor has established a positive (eg, cooperative) reputation. Nonetheless, individuals do not know about all interactions among actors in reality since tracing back to all relational interactions of the other actor(s) is unfeasible (Milinski 2016). Instead, people may rely on status-based characteristics to approximate someone's trustworthiness based on their social status and related performance expectations. That is, if social images exist in a population where greater competences and trustworthiness are associated with certain groups (eg, native population) in relation to lower social status (eg, refugees), people may infer a positive/negative reputation based on strangers' apparent status so that the level of trust can differ in favour of trustees perceived as higher status. Relatedly, Zhang and colleagues (2019), through a field experiment demonstrate that native people are more likely to offer help (being more prosocial) to high-status immigrants than low-status immigrants, characterised based on their language skills.

Drawing on the status characteristics theory (Ridgeway 1991), I focus on a few points where the effect of reputational information on trusting and trustworthy behaviour in *one-shot* and *indirectly observable social dilemma settings* may be confounded by social status, thereby influencing expectations of vulnerability to one's actions in cooperation. In explaining this confounding bias, I delve into the *assumption of excludability* in the conventional causal path of cooperation through reputational information, which ensures that exogenous variation created by a random assignment influences the outcome through only the variable of interest (eg, availability of reputational information). Otherwise, the estimated treatment effect cannot be regarded as causal. In particular, the treatment assignments using the observability or availability of reputational information in different contexts (Chu et al 2019; Resnick et al 2006; Yoeli et al 2013) do not only lead to belief updating on reputations to understand the variation in outcomes (eg, cooperation for public goods provision) but also affect people's prior beliefs interacted with various omitted factors. For instance, Yoeli et al (2013) collaborated with a utility company for a demand response programme to prevent blackouts because of excessive use of electricity. The demand response programme requires residents to voluntarily sign up to authorise the company to restrict their central use of air conditioners via a remote switch-in-line during peak hours. Participation in this programme by residents in the city is socially optimal because the inconvenience is minimal for households, compared to the societal cost of blackouts. In doing so, the main experimental treatment was to randomly vary whether residents' neighbours could tell who signed up for the programme, looking at publicly available sheets in a communal area near their home. Their study found that making participation in a public good observable by neighbours in a field experiment renders people three times more likely to cooperate. Nonetheless, this conventional design may not strictly hold the *assumption of excludability.*

More precisely, the excludability violation in this design may occur because people's previous interactions with neighbours are entangled with their social status in the community. That is, random assignment to treatment might also be triggering other possible causes to the outcome. For example, if neighbour$_i$ (high-status) is aware of the social status of the neighbour$_j$ (low-status), the reputational information on neighbour$_j$'s choice to participate in a collective action could influence neighbour$_i$'s vulnerability updating through neighbour$_i$'s expectations about how the community perceives the social status of neighbour$_j$ rather than only the reputational information itself. Even if the reputational information on neighbour$_j$ is positive, lower expectations on neighbour$_j$'s contribution due to the social status of neighbour$_j$ may still lead neighbour$_i$ to remain vulnerable to the defection of neighbour$_j$. In addition, neighbour$_i$ (high-status) is vulnerable to incurring reputational cost if they choose not to contribute to public goods since the community primarily expects high-status neighbours to contribute.

In elaborating on the excludability violation, one's expectation of vulnerability to others' free-riding behaviour is not only affected by how most people perceive the position and characteristics of one's social status but also one's identity[2] (eg, political identity). That is, if neighbour$_i$ were informed that the specific collective action at hand is coordinated by a specific third party (eg, a political party) or neighbour$_i$ was aware of the political identity of neighbour$_j$, it is expected that neighbour$_i$ would question the trustworthiness of the given reputational information and update their expectation of vulnerability to neighbour$_j$'s action accordingly. Together, social norms are not homogeneous across community or even different political identities in the same community (Ellingsen et al 2013). For instance, previous research showed that political conservatives are more likely to refuse the availability of reputational information in the context of energy conservation (Costa and Kahn 2013).

Overall, the direct effect of the availability of reputational information may boil down to the extent to which the information on actors (eg, their social status and identity) is incomplete among individuals in a certain social dilemma. For example, Yoeli et al (2013) found that homeowners living in the apartments are significantly influenced by the availability of reputational information on whether their neighbours participated in the energy-saving programme, while renters are not responsive to the same indirect reciprocity to contribute to the public good. Researchers interpreted that renters are more temporary than homeowners and, therefore, less likely to invest in long-term relationships with other neighbours. If I am allowed to counterfactually assert that if the social status of renters and homeowners were visible across the participants in the same building, then we would be able to better comprehend the limitation with the assumption of excludability in the study at hand. In line with the status characteristic theory, the renters (eg, low-status) would be expected to contribute less to the public good compared to their homeowners because of their hierarchical status by which higher-status members (eg, homeowners) are expected to be more influential in participation (Ridgeway 1991), regardless of the duration of interactions. Thus, it is a strong assumption that exposure to reputational information directly causes cooperation in large-scale collective actions. Rather, I emphasise that the effect of reputational information becomes heavily contingent on status characteristics of individuals in collective actions. Individuals in different social status positions have distinct expectations of being vulnerable to untrustworthy or distrusting action. In contrast

[2] In line with the social identity theory (Tajfel et al 1971), trust can also be influenced by identity-related characteristics of a trustor and a trustee, as well as status-based characteristics. For instance, if a trustee and a trustor share the same salient social categories, it is expected that the level of trust will be biased towards in-group members and biased against out-group members.

to field experiments, the laboratory setting enables researchers to anonymise the participants regarding their social cues in the Trust game to better isolate the effect itself, yet the structure of the game still holds the fact that if the trustee is aware that the trustor can give no opportunity to entail any possibility of betrayal (not sending any amount), trustee's expectation of being vulnerable to the trustor is contingent on their position with respect to the hierarchy of the game.

IV. Reputation, Vulnerability, and Mistrusted Communities

Given the aforementioned theoretical reconsiderations, I highlight the potential challenges for researchers and policymakers in social justice and refugee studies to disentangle how reputational information may backfire or help refugees and host populations to cooperate. Refugees are often perceived and represented as 'economic burden', 'culturally incompatible' or 'mistrusting for security reasons' in the public sphere through media or political institutions (Dempster & Hargrave 2017). Such crystallised stereotypes may operate as cognitive shortcuts[3] (cue-taking) in signalling a social status or social identities that undermine the positive impact of reputational information on mistrusted communities when host populations engage in trustworthy or trusting behaviour with them. That is, if the host population receives communications through media channels about positive contributions of refugees to their host communities (eg, helping the rejuvenation of local parks), this reputational information contradicting the expectations on refugees based on their social status characteristics may not entirely update the host population's expectations of vulnerability to refugees' actions in trusting relationships since most perceive low-status groups to be less trustworthy (Ridgeway, 1997). In addition to this challenge, previous research also shows that when people are exposed to negative reputational information, this strengthens their existing biases if the given reputation does not contradict their prior expectation (*confirmatory bias*) (Chu et al 2019). This may explain how the host population polarises their expectations about refugee communities over time when they receive negative reputational information on refugees. The other side of the coin is that refugees are in a certain social position to expect trusting behaviour of the host communities with a greater degree of vulnerability in the first place. This asymmetry is akin to the structure of the binary Trust game where the trustee cannot take any action to challenge the non-cooperative trustor, who provides no opportunity to entail a possibility of betrayal.

[3] I refer to the bounded rationality account of cue receptivity, which assumes that people avoid costly cognitive efforts for belief formation. Instead, they base their expectations on value-consistent positions (eg, social status) (Bakker et al 2019).

In light of the aforementioned discussion, let us consider an asylum seeker who hopes for cooperation to grant protection in another country–*impersonal trust* arises here. This refers to an indirect trusting relationship, in which trusting action is mediated by an agent or structure (Shapiro 1987). Let's assume that there is no binding or functional international treaty to urge countries to accept each asylum seeker. Impersonal trust is mostly perceived as shared norms rather than an individually calculated phenomenon (Hosmer 1995) and yields a humanitarian expectation that others will trust asylum seekers. However, host populations have a structural opportunity to move first to decide on whether to trust or not to welcome asylum seekers into the country.[4] This influences the vulnerability of asylum seekers to the betrayal of impersonal trust, while the countries or host populations build up their trusting relationships with other populations and countries, relying on interdependence, familiarity and known reputation with whom they have had past interactions rather than one-shot interactions (Granovetter 1985). That said, an asylum seeker who is not trusted may incur reputational harm against the host populations through international or domestic humanitarian organisations. Hendrix and Wong (2013), for instance, showed that operating reputational concerns by international human rights organisations could explain larger effects on altering human rights abusing behaviour in autocracies. Hence, the larger action space of asylum seekers in an impersonal trusting relationship with host populations may change the degree of vulnerability of the host population in cooperation in spite of status hierarchies between host populations and asylum seekers.

After discussing the challenges for asylum seekers in their initial one-off impersonal trusting relationship with host populations, let us continue with their probationary period[5] if refugee status is granted. Theoretically, trust and reciprocity can lead to sustainable cooperation when an interaction is repeated among strangers in interpersonal relationships if they particularly think that they may encounter each other in the future (Bicchieri et al 2004). Nonetheless, in many circumstances, keeping track of repeated interactions to develop a rationally optimal strategy in trusting relationships among actors is cognitively exhaustive and costly. Alternatively, researchers have therefore drawn attention to individual learning theories (Buskens and Raub 2002; Skyrms 1990) or how disclosing reputational information may serve as a form of forging trust and cooperation through reciprocal *self-closure*, thereby making themselves willingly vulnerable to each other (Feinberg et al 2012; Kopányi-Peuker et al 2017; Sommerfeld et al 2007; 2008; and many others). Given these considerations, I choose to engage in how the availability of reputational information may enhance the vulnerability of refugees

[4] Asylum seeking applicants will not be granted asylum if the government believes that the applicant may render the country vulnerable to potential threats (Home Office 2019).

[5] For instance, after the initial asylum application, an applicant can be recognised as a refugee or granted asylum with five years' leave to remain in the UK in order to be able to apply later for indefinite leave to remain (Home Office, 2019). This process is defined as a transition/interim period.

(low-status) and host populations (high-status) to each other in inducing impersonal trust and cooperation through counterfactual examples.

Let us imagine a refugee (a former political prisoner in country Y) who has just gained initial trust from the host country X to resettle. Once refugee status is granted by the host country, the refugee as a trustee is expected to submit confidential personal information and evidence to the host country X, even though discouraged to disclose such information to strategically minimise the likelihood of betrayal and exploitation in trusting relationships. In sum, the host country X is only vulnerable to the degree where reputational information provided by the refugee is inaccurate, thereby abusing the trust. By contrast, the refugee enters into a social contract–it is controversial to define whether this contract is voluntarily or forced–accepting a higher level of vulnerability. In this case, to sustain cooperation, the refugee is obliged to reciprocate given trust, taking a leap in the dark without the possibility to enhance an impact on the host country's vulnerability to their trusting relationship. In addition to this structural asymmetry in self-closure, refugees are also represented as low-status in social hierarchies. This situation strongly signals that the wider population expects refugees to be less trustworthy vis-à-vis sustaining cooperation. Taken together, the asymmetric enhanced vulnerability and social status of the trustee in the trusting relationship may inevitably underpin the expectation that the capacity of refugees' trustworthiness is undermined by the culture of suspicion towards them (Hynes 2009).

Let us reconsider the same refugee, holding the characteristics of the host country X and time constant, with regular assistance by international or domestic human rights organisations throughout the resettlement period. Instead of exposing their reputational information to the host country rapidly, these organisations can act as the guardian of trust (a third party) in order to respect what and when the refugee is willing to disclose information and attempt to build trust gradually. In sharing reputational information with the organisations, the refugee, in turn, is given an opportunity to spread some reputational information about their host country to public spheres where companies, customers and tourists, for instance, make their humanitarianly conscious decisions with the host country (eg, cultural boycotts). Note that the role of these organisations is not to punish norm violation but to coordinate norm enforcement, assisting refugees to communicate reputational information on the host country to the wider public spheres. As outlined earlier, if the trustor becomes aware of the possibility that the trusted actor (refugee) has an opportunity to incur reputational harm, the trustor may expect less trustworthiness because of their increased vulnerability to the actions of the trustees in one-off interactions. In contrast, the impersonal trust relationship coordinated through a third party described above entails a variant where both refugees and host countries may have the possibility to build trust by conditioning their enhanced vulnerability on the other's enhanced vulnerability in small steps, thereby transforming themselves into successful cooperators. Therefore, given that refugees are trapped in their lower social position and mistrusting status characteristics when building trust relationships with such mistrusted communities, host populations need to recognise that enhancing their vulnerability towards refugees may help alleviate lower expectations on the trustworthiness of refugees in cooperation.

V. Implications and Conclusion

The implications of this chapter centre around that (1) the conventional frameworks in understanding social dilemmas undertheorise the endogeneity of social status and structural vulnerability in trusting and trustworthy behaviours, drawing attention to the limited action space of trustees; (2) the enlarged action space of asylum seekers in host countries might resolve free-riding in cooperation if the host country's vulnerability to the actions of asylum seekers is enhanced; (3) refugees and host countries may have the possibility to build trust by conditioning one's own enhanced vulnerability through reputational information on the other's enhanced vulnerability in small steps rather than in leaps and bounds; (4) the exposure to or communicating reputational information in promoting cooperation across individuals may be confounded by social status and identities to sustain cooperation when someone's social status or identities carry contrary evidence against provided reputational information.

First, these implications highlight that further research in behavioural social sciences should experimentally examine how the enlargement of action space for trustees in different social dilemmas influences trustor's expectation about their vulnerability to trustees in both one-off and repeated interactions. In doing so, further studies can also measure trustee's beliefs about trustor's vulnerability and trustee's beliefs about trustor's guesses of the trustee's vulnerability. This would help us better interpret to what extent expectations on vulnerabilities affect trusting and trustworthy behaviour. Second, policymakers in refugee-hosting countries responsible for social justice should keep in mind that asylum seekers may become more cooperative in application processes if the rapport with them relies on a gradual mechanism where reciprocal vulnerabilities develop in small steps rather than stigmatising asylum seekers with suspicion in the first place. In this regard, the first interactions with asylum seekers can be conducted by professional refugee care providers with whom they can regain a sense of security. Third, a clear causal identification and designing interventions are important when researchers deal with certain vulnerable groups. To my knowledge, the majority of studies using reputational information have not unveiled the role of vulnerability in understanding how expectations of being vulnerable to one's action help or undermine cooperation. Relevantly, in future studies, scholars should account for heterogonous characteristics of people's social status moderating the effect of reputational information. That is, communicating positive reputational information on the social image of low-status communities in collective actions may not simply lead people to update their expectations on the trustworthiness of those communities.

Therefore, one approach to improve the causal identification of reputational information in solving social dilemmas is to curb the impact of people's social status by either statistical controls or ideally manipulating the perceived social status of people. Overall, I hope this chapter clarifies some issues on the causal identification caveats in dealing with the endogeneity of vulnerability through trusting and trustworthy behaviour and helps researchers and policymakers clearly understand how they can improve their interventions to facilitate cooperation in relevant social dilemmas.

References

Ananish, Chaudhuri and Lata, Gangadharant. (2007). 'An Experimental Analysis of Trust and Trustworthiness.' *Southern Economic Journal* 27.

Bacharach, M, Guerra, G and Zizzo, D J (2007). 'The Self-fulfilling Property of Trust: An Experimental Study.' *Theory and Decision* 63 (4): 349–88.

Bakker, B, Lelkes, Y and Malka, A (2019). *Understanding Partisan Cue Receptivity: Tests of Predictions from the Bounded Rationality and Expressive Utility Perspectives*.

Ben-Ner, A and Putterman, L (2009). 'Trust, Communication and Contracts: An Experiment.' *Journal of Economic Behavior & Organization*, 70 (1–2): 106–21.

Berg, J, Dickhaut, J and McCabe, K (1995). 'Trust, Reciprocity, and Social History.' *Games and Economic Behavior* 10 (1): 122–42.

Bicchieri, C, Duffy, J and Tolle, G (2004). 'Trust Among Strangers.' *Philosophy of Science* 71 (3): 286–319.

Bowles, S (2008). 'Policies Designed for Self-interested Citizens may Undermine "the Moral Sentiments": Evidence from Economic Experiments.' *Science* 320 (5883): 1605–09.

Buskens, V and Raub, W (2002). 'Embedded Trust: Control and Learning.' *Advances in Group Processes*, 19: 167–202.

Chaudhuri, A (2011). 'Sustaining Cooperation in Laboratory Public Goods Experiments: A Selective Survey of the Literature.' *Experimental Economics* 14 (1): 47–83.

Chu, J, Li, G, Loyalka, P, Liu, C, Rosa, L & Li, Y (2019). 'Stuck in Place? A Field Experiment on the Effects of Reputational Information on Student Evaluations.' *Social Forces*.

Coleman, J S (1994). *Foundations of social theory*. Harvard University Press.

Colquitt, J A, Scott, B A and LePine, J A.(2007). 'Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance.' *Journal of Applied Psychology* 92 (4): 909.

Costa, D L and Kahn, M E (2013). 'Energy Conservation "Nudges" and Environmentalist Ideology: Evidence from a Randomized Residential Electricity Field Experiment.' *Journal of the European Economic Association* 11 (3): 680–702.

Dempster, H and Hargrave, K (2017). 'Understanding Public Attitudes Towards Refugees and Migrants.' *London: Overseas Development Institute & Chatham House*.

Deutsch, M (1962). *Cooperation and Trust: Some Theoretical Notes.*

Diekmann, A, Jann, B, Przepiorka, W and Wehrli, S (2014). 'Reputation Formation and the Evolution of Cooperation in Anonymous Online Markets.' *American Sociological Review* 79 (1): 65–85.

Dirks, K T and Ferrin, D L (2002). 'Trust in Leadership: Meta-nalytic Findings and Implications for Research and Practice.' *Journal of Applied Psychology* 87 (4): 611.

Ellingsen, T, Herrmann, B, Nowak, M, Rand, D G and Tarnita, C (2013). 'Civic Capital in Two Cultures: The Nature of Cooperation in Romania and USA.' *Available at SSRN 2179575*.

Ermisch, J, Gambetta, D, Laurie, H, Siedler, T and Noah Uhrig, S C (2009). 'Measuring People's Trust.' *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172 (4): 749–69.

Falk, A and Zehnder, C (2013). 'A City-Wide Experiment on Trust discrimination.' *Journal of Public Economics* 100: 15–27.

Fehr, E, Fischbacher, U, Von Rosenbladt, B, Schupp, J and Wagner, G G (2003). *A Nation-Wide laboratory: Examining Trust and Trustworthiness by Integrating Behavioral Experiments into Representative Survey*.

Fehr, E and Gachter, S (2000). 'Cooperation and Punishment in Public Goods Experiments.' *American Economic Review* 90 (4): 980–94.

Fehr, E and Gächter, S (2002). 'Altruistic Punishment in Humans.' *Nature* 415 (6868): 137–40.

Fehr, E and Rockenbach, B (2003). 'Detrimental Effects of Sanctions on Human Altruism.' *Nature* 422 (6928): 137–40.

Feinberg, M, Willer, R, Stellar, J and Keltner, D (2012). 'The Virtues of Gossip: Reputational Information Sharing as Prosocial Behavior.' *Journal of Personality and Social Psychology* 102 (5): 1015.

Fischbacher, U and Gachter, S (2010). 'Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments.' *American Economic Review* 100 (1): 541–56.

Gächter, S, Herrmann, B and Thöni, C (2004). 'Trust, Voluntary Cooperation, and Socio-Economic Background: Survey and Experimental Evidence.' *Journal of Economic Behavior & Organization* 55 (4): 505–31.

Gambetta, D. (1988). *Trust: Making and Breaking Cooperative Relations*.

Glanville, J L (2004). 'Voluntary Associations and Social Network Structure: Why Organizational Location and Type are Important.' *Sociological Forum* 19 (3): 465–91.

Glanville, J L and Paxton, P. (2007). 'How Do We Learn to Trust? A Confirmatory Tetrad Analysis of the Sources of Generalized Trust.' *Social Psychology Quarterly* 70 (3): 230–42.

Granovetter, M (1985). 'Economic Action and Social Structure: The Problem of Embeddedness.' *American Journal of Sociology* 91 (3): 481–510.

Hendrix, C S and Wong, W H (2013). 'When is the Pen Truly Mighty? Regime Type and the Efficacy of Naming and Shaming in Curbing Human Rights Abuses.' *British Journal of Political Science* 43 (3): 651–72.

Ho, T-H and Weigelt, K (2005). 'Trust Building Among Strangers.' *Management Science* 51 (4): 519–30.

Homans, G C (1958). 'Social Behaviour as Exchange.' *American Journal of Sociology*.

Home Office. (2019). *User Guide to Home Office Immigration Statistics*. 133.

Hosmer, L T (1995). 'Trust: The Connecting Link Between Organizational Theory and Philosophical Ethics.' *Academy of Management Review* 20 (2): 379–403.

Hynes, P (2009). 'Contemporary Compulsory Dispersal and the Absence of Space for the Restoration of Trust.' *Journal of Refugee Studies* 22 (1): 97–121.

Lo Iacono, S and Sonmez, B (2021). 'The Effect of Trusting and Trustworthy Environments on the Provision of Public Goods.' *European Sociological Review* 37 February (1):155–68.

Jordan, J J, Hoffman, M, Nowak, M A and Rand, D G (2016). 'Uncalculating Cooperation is Used to Signal Trustworthiness.' *Proceedings of the National Academy of Sciences* 113 (31): 8658–63.

Jordan, J, Peysakhovich, A and Rand, D G (2015). 'Why We Cooperate.' *The Moral Brain: A Multidisciplinary Perspective* 87.

Keser, C and Özgümüs, A (2018). 'Trust, Vulnerability and Trustworthiness.' *Economics Letters* 163: 149–51.

Kiyonari, T, Yamagishi, T, Cook, K S and Cheshire, C (2006). 'Does Trust Beget Trustworthiness? Trust and Trustworthiness in Two Games and Two Cultures: A Research Note.' *Social Psychology Quarterly* 69 (3): 270–83.

Kocher, M G, Martinsson, P, Matzat, D and Wollbrant, C (2015). 'The Role of Beliefs, Trust, and Risk in Contributions to a Public Good.' *Journal of Economic Psychology* 51: 236–44.

Kollock, P (1998). 'Social Dilemmas: The Anatomy of Cooperation.' *Annual Review of Sociology* 24 (1): 183–214.

Kopányi-Peuker, A, Offerman, T and Sloof, R (2017). 'Fostering Cooperation Through the Enhancement of own Vulnerability.' *Games and Economic Behavior*, 101: 273–90.

Kreps, D M (1990). *Game Theory and Economic Modelling*. Oxford University Press.

Kuwabara, K. (2015). 'Do Reputation Systems Undermine Trust? Divergent Effects of Enforcement Type on Generalized Trust and Trustworthiness.' *American Journal of Sociology* 120 (5): 1390–428.

Ledyard, O (1995). 'Public Goods: Some Experimental Results.' *Handbook of Experimental Economics* 1.

List, J A (2007). 'On the Interpretation of Giving in Dictator Games.' *Journal of Political Economy* 115 (3): 482–93.

Luhmann, N (2018). *Trust and Power*. John Wiley & Sons.

Mayer, R C, Davis, J H and Schoorman, F D (1995). 'An Integrative Model of Organizational Trust.' *Academy of Management Review*, 20 (3): 709–34.

McAllister, D J (1995). 'Affect-and Cognition-Based Trust as Foundations for Interpersonal Cooperation in Organizations.' *Academy of Management Journal* 38 (1): 24–59.

Milinski, M (2016). 'Reputation, a Universal Currency for Human Social Interactions.' *Philosophical Transactions of the Royal Society B: Biological Sciences* 371 (1687): 20150100.

Möllering, G (2006). 'Trust Beyond Risk: The Leap of Faith.' *Trust: Reason, Routine, Reflexivity* 105–26.

Montgomery, K, Jordens, C F C and Little, M (2008). 'How Vulnerability and Trust Interact During Extreme Events: Insights for Human Service Agencies and Organizations.' *Administration & Society* 40 (6): 621–44.

Mulder, L B, Van Dijk, E, De Cremer, D and Wilke, H A (2006). 'Undermining Trust and Cooperation: The paradox of Sanctioning Systems in Social Dilemmas.' *Journal of Experimental Social Psychology* 42 (2): 147–162.

Olson, M (2012). 'The Logic of Collective Action' [1965]. *Contemporary Sociological Theory*, 124.

Przepiorka, W and Berger, J (2017). 'Signalling Theory Evolving: Signals and Signs of Trustworthiness in Social Exchange.' *Social Dilemmas, Institutions and the Evolution of Cooperation. Berlin: De Gruyter Oldenbourg* 373–92.

Przepiorka, W, Norbutas, L and Corten, R (2017). 'Order without law: Reputation Promotes Cooperation in a Cryptomarket for Illegal Drugs.' *European Sociological Review* 33 (6): 752–64.

Resnick, P, Zeckhauser, R, Swanson, J and Lockwood, K (2006). 'The Value of Reputation on eBay: A Controlled Experiment.' *Experimental Economics* 9 (2): 79–101.

Rigdon, M L, McCabe, K A and Smith, V L (2007). 'Sustaining Cooperation in Trust Games.' *The Economic Journal* 117 (522): 991–1007.

Ridgeway, C L (1978). 'Conformity, Group-Oriented Motivation, and Status Attainment in Small Groups.' *Social Psychology* 175–88.

Ridgeway, C L (1982). Status in groups: 'The Importance of Motivation.' *American Sociological Review* 76–88.

Ridgeway, C L and Berger, J (1986). 'Expectations, Legitimation, and Dominance Behavior in Task Groups.' *American Sociological Review* 603–17.

Ridgeway, Cecilia L (1991). 'The Social Construction of Status Value: Gender and Other Nominal Characteristics.' Social Forces 70 (2): 367–86.

Ridgeway, C L (1997). 'Where Do Status Value Beliefs Come From? New Developments.' *Status, Network, and Structure* 137–58.

Robertson, R, Gockel, C and Brauner, E (2013). 'Trust your Teammates or Bosses? Differential Effects of Trust on Transactive Memory, Job Satisfaction, and Performance.' *Employee Relations*.

Romano, A, Balliet, D, Yamagishi, T and Liu, J H (2017). 'Parochial Trust and Cooperation Across 17 Societies.' *Proceedings of the National Academy of Sciences* 114 (48): 12702–07.

Rothstein, B (2000). 'Trust, Social Dilemmas and Collective Memories.' *Journal of Theoretical Politics* 12 (4): 477–501.

Rousseau, D M, Sitkin, S B, Burt, R S and Camerer, C (1998). 'Not so Different After All: A Cross-Discipline View of Trust.' *Academy of Management Review* 23 (3): 393–404.

Shapiro, S P (1987). 'The Social Control of Impersonal Trust.' *American Journal of Sociology* 93 (3): 62–658.

Simmel, G (1950). 'The Sociology of Georg Simmel. Transl., ed and intr. by K H Wolff.' New York: Free Press.

Simpson, B and Eriksson, K (2009). 'The Dynamics of Contracts and Generalized Trustworthiness.' *Rationality and Society* 21 (1): 59–80.

Skyrms, B (1990). *The Dynamics of Rational Deliberation*. Harvard University Press.

Sommerfeld, R D, Krambeck, H-J and Milinski, M (2008). 'Multiple Gossip Statements and their Effect on Reputation and Trustworthiness.' *Proceedings of the Royal Society B: Biological Sciences* 275 (1650): 2529–36.

Sommerfeld, R D, Krambeck, H-J, Semmann, D and Milinski, M (2007). 'Gossip as an Alternative for Direct Observation in Games of Indirect Reciprocity.' *Proceedings of the National Academy of Sciences* 104 (44): 17435–40.

Tajfel, H, Billig, M G, Bundy, R P and Flament, C (1971). 'Social Categorization and Intergroup Behaviour.' *European Journal of Social Psychology* 1 (2): 149–78.

Trifiletti, E and Capozza, D (2011). 'Examining Group-Based Trust with the Investment Game.' *Social Behavior and Personality: An International Journal* 39 (3): 405–09.

Uslaner, E M (2002). *The Moral Foundations of Trust*. Cambridge University Press.

Van Lange, P A (1999). 'The Pursuit of Joint Outcomes and Equality in Outcomes: An Integrative Model of Social Value Orientation.' *Journal of Personality and Social Psychology* 77 (2): 337.

Willer, R (2009). 'Groups Reward Individual Sacrifice: The Status Solution to the Collective Action Problem.' *American Sociological Review* 74 (1): 23–43.

Yamagishi, T, Matsuda, M, Yoshikai, N, Takahashi, H & Usui, Y (2009). 'Solving the Lemons Problem with Reputation.' *ETrust: Forming Relationships in the Online World* 73–109.

Yoeli, E, Hoffman, M., Rand, D G and Nowak, M A (2013). 'Powering up with Indirect Reciprocity in a Large-Scale Field Experiment.' *Proceedings of the National Academy of Sciences* 110 (Supplement 2), 10424–29.

Zhang, N, Aidenberger, A, Rauhut, H and Winter, F (2019). 'Prosocial Behaviour in Interethnic Encounters: Evidence from a Field Experiment with High-and Low-Status Immigrants.' *European Sociological Review* 35 (4): 582–97.