Trust, Implicit Attitudes, and the Malleability of Group Identities

Sarah Songhorian

1. *Introduction*

The concept of trust, just as many concepts we ordinarily use in our moral, social, and political discourse, is a complex and multifaceted one (McLeod 2015; Hawley 2012; Simpson 2012; Baier 1986). By applying it to a variety of different contexts, it is hard to have a good sense of its conceptual boundaries and of whether using it is appropriate or not in a given context. Whether, for instance, we are actually talking about the same concept when we consider self-trust, interpersonal trust, or trust in institutions is an open and difficult question to be asked. Furthermore, there is a branch of empirical research that is growing quite fast on the impact implicit attitudes and biases have on modulating trust.

The aim of this paper is underlining the necessity for a more fine-grained definition of our ordinary conception of trust that, in line with the data I will discuss, focuses on the need to distinguish clearly two different levels within the concept of trust itself. As I will show, this will also help us to get a better understanding of the role trust can and should play in ethics. Before moving to the consideration of the empirical data and to the revision of the concept of trust I will suggest, it is important to get a sense of what we ordinarily mean by trust.

Pretheoretically, trust is an important force prompting us to rely on others – human being, institutions, or authorities – and to build a relationship with them. While this general understanding of trust can be applied to all its forms and varieties, in what follows I will restrict myself to the consideration of interpersonal trust as it is the first one to be modulated by the data I will consider. This, however, does not mean that implicit attitudes

DOI: 10.4454/teoria.v39i1.61

cannot have an impact also on other forms of trust or that this possible outcome is irrelevant, but simply that it is secondary.

Basically, A trusts B only if A relies upon B to meet her commitments – whether by doing something, by saying something, by behaving in a certain way, or by being in a certain way – and, thus, A *believes* B to be trustworthy (analogously, Hawley 2012: 6). It has to be noticed, however, that the fact that A (the trustor) believes B (the trustee) to be trustworthy does not imply that B *is* actually trustworthy. Therefore, A's trust can be unwarranted or ungrounded since it may target a non-trustworthy subject as if she were to be trusted (McLeod 2015). This feature of our colloquial understanding of trust is usually easily identified and it is because of it that trust is considered important in our interpersonal relationships but at the same time risky, since it implies that subjects rely on others for matters that are of interest to them and that those others may not deserve to be considered reliable.

This ordinary conception of trust is often taken as an obvious requirement for cooperation, and, being the latter of interest for morality, it is also taken as a requirement for morality to flourish (Baier 1986: 232). However, this does not *per se* imply that every time we trust someone we are in a moral relationship to that person. Trust can, in fact, also refer to amoral situations or interactions (as we shall see in § 2 and 3).

The data I will review in what follows (§ 2 and 3) will show another, deeper, reason to think that trust can be risky. While the possibility of trust being unwarranted or ungrounded is recognized by anyone who has a concept of trust – since the most common reason to feel one's trust betrayed is the non-correspondence between one's belief that the other is trustworthy and the other's actual trustworthiness –, the risk of implicit modulation is hardly recognized or considered. Very few people would accept that their trusting attitudes are malleable to several unconscious drives. In fact, while we often are aware of our decisions and actions taking place in a certain context, most of us are unconscious of the very *influence* such situational factors can play (Herdova 2016: 52). In order to understand exactly what aspect of trust – or what level – is modulated by these drives, I will have to provide a two-level conceptualization of trust (§ 4).

In this paper, I will restrict myself in two respects. On the one hand, I will only consider one specific unconscious drive that modulate our moral interaction with others, among the many identified in the literature – namely, group identity. Hence, I will not be interested in other well-known springs of emotional and situational modulation such as, for instance, the fact that a clean setting may decrease the severity of our moral judgments (Huang 2014; Schnall *et al.* 2008) or that a good scent may promote reciprocity and charity (Liljenquist 2010). While these data could be used to question our colloquial understanding of trust as a stable attitude, just as much as situationism has used them to challenge the existence of stable dispositions and personality traits (e.g. Harman 1999, 2000, 2001, 2003, 2009; Doris 1998, 2002, 2010; Appiah 2008), it is not my aim in this work to follow that path. The analysis I will put forth aims at showing how the concept of trust itself requires a better conceptualization rather than at suggesting that individuals' trusting traits are unstable (which most likely are). On the other hand, I will focus only on the effect of these drives on interpersonal trust (henceforth, trust *simpliciter*). This, however, does not imply that these unconscious drives can only impact the extent to which we trust others in face-to-face interactions. Quite the contrary, they can also increase or decrease the extent to which we empathize with them (Xu *et al.* 2009), the extent to which we behave altruistically (see for instance the research on parochial altruism; Bernhard *et al.* 2006) or cooperatively (Greene 2013), and the extent to which we help others (Levine *et al.* 2005); just as much as they may have a secondary impact on other forms of trust as well. And yet, the focus on possible influences on interpersonal trust is motivated by at least two reasons. First, since trust is an attitude and not an actual behaviour, it might be that altruism and cooperation depend at least in part on our trusting attitudes and not the other way around (Baier 1986: 232 considers obvious the connection between cooperation and trust). Hence, seeing the impact of these unconscious drives on trust can reveal a more fundamental effect as opposed to focusing on behaviour. Second, trusting attitudes are easy to test and the data on them are quite clear once exposed.

2. *Group Entitativity and Social Categorizing*

As Joshua Greene points out in his *Moral Tribes*, «morality evolved to enable cooperation» (2013: 23). According to Greene, cooperation is crucial for morality. For the purposes of this work, it is important to underline how cooperation is rendered possible by trusting those with whom we cooperate. Hence, it does not seem exaggerated to claim that, in order to have cooperation, we need to have some level of trust (Baier 1986: 232). Trust is, therefore, taken to be necessary for cooperation, even though it may not be sufficient.

Greene also adds, that:

> Biologically speaking, humans were designed for cooperation, *but only with some people*. Our moral brains evolved for cooperation *within groups*, and perhaps only within the context of personal relationships. Our moral brains did not evolve for cooperation *between groups* (at least not *all* groups) (Greene 2013: 23, emphasis in original).

This insight into how our ability to cooperate is limited by our group affiliation depends on several different data. First, there is ethological evidence on animals cooperating with conspecifics but not with other species. Even more interestingly, within the same species, it is more common to see animals cooperating with kin rather than with non-kin (Hamilton 1964; Wynne-Edwards 1962; Kropotkin 1908). Second, psychological research on social categorizing has shown in the past 40 years that the behavioural influences of our recognition of a group identity are very solid. These influences can either be explicit or implicit depending on whether the subject is or is not aware of their functioning and whether she endorses, avows and self-attributes them or not (Levy 2017: 3). The more one perceives group entitativity – that is, the more one perceives a group as a real entity characterized by similarity and cohesion among its members (Plötner *et al.* 2016; Dasgupta *et al.* 1999; Hamilton *et al.* 1998; Yzerbyt *et al.* 1998) –, the more she would be likely to favour her own group members over people belonging to other groups. To detail this second line of research, I will briefly expose some of the data collected to show how cooperation and trust are implicitly modulated by social categorizing both in adults and in children. The following discussion will be grounded on the assumption, shared by Baier and Greene among others, of an obvious or at least commonly recognized relationship between cooperation and trust. While this assumption most certainly deserves a deeper consideration and calls for an argument in its favour based on independent grounds, in this context I will have to take it for granted for the sake of the argument.

When adults have to predict whether they will receive more money from an unknown allocator belonging to their same group as opposed to one belonging to another group, they strongly prefer trusting their in-group members (from 76% to 89%, see Foddy *et al.* 2009: 421), if they are told that the allocator knows about their own group membership (*common-knowledge condition*). If this last condition is not met – i.e. when participants know that the allocator is unaware of their group membership –, what matters is the stereotype associated with both in-groups and out-groups (*private-*

knowledge condition). Foddy and colleagues' (2009) subjects were economic and nursing majors. Hence, when membership was known by all participants, subjects trusted their in-groups more (or, in other words, thought ingroup members to be more trustworthy); whereas when the group affiliation was clearly unknown to the allocator, subjects decided whether to trust others or not making the stereotypes associated with economic and nursing majors much more salient: «The percentage of participants who chose an ingroup allocator was larger when the out-group was economics majors (80%) than when it was nursing majors (41%)» (Foddy *et al.* 2009: 421). To control whether subjects will still prefer trusting in-groups as opposed to out-groups when they could choose a sure thing (AU$ 6.00 from the experimenter) and avoid trusting altogether, Platow and colleagues conducted two further studies (Platow *et al.* 2012). The results of these studies were in line with the data collected by Foddy and colleagues (2009): participants decided to trust in-groups even when they could have dropped out in the common-knowledge condition. Hence, the data on trusting attitudes towards ingroup members were enhanced by these further researches: subjects do not only trust in-group over out-group when they had to trust someone (*relative trust*), but also when they had the opportunity to opt out (*absolute trust*).

Similar data on adults were also collected in several studies on the investment game (Stanley *et al.* 2011; Güth *et al.* 2008; Tanis, Postmes 2005; DeBruine 2002). The investment game – also known as the trust game – is an economic game often used to measure the extent to which people trust others (Johnson, Mislin 2011; Berg *et al.* 1995). In this game, subjects have to decide whether to invest the money they have earned by participating in the experiment. Once the money is invested, experimenters tell participants that the money will be given augmented – e.g. tripled in the studies conducted by Tanis and Postmes (2005) and by Güth and colleagues (2008); and quadrupled in the one by Stanley and colleagues (2011) – to another individual, who can choose whether to reciprocate or not. It is in the participants' best interest to invest as much money as possible if they trust the counterpart. In this kind of research, «the measure of trust was an ecologically relevant consequential decision about how much money to risk in each interaction» (Stanley *et al.* 2011: 7712), rather than an explicit assessment by the participants of how much they trust others or of whether they thought the counterpart was actually trustworthy. This is extremely important for at least two reasons. First, given that the presence of trust is inferred from monetary interactions, the authors can grant that the participants' actual and conscious motivation is the desire of gaining money (i.e. it is self-interest that moves them rather than a benevolent or altruistic drive). However, in order for the subjects to gain the most money they could, they had to actually trust their counterparts. Therefore, despite participants' motivation could be taken to be mere self-interest, the authors can easily claim that the latter has to be backed with trust for the subjects to actually behave as they do: if they were not to trust the counterpart to send them an adequate amount of money back, they would decide not to invest at all precisely for their own self-interest. Second, as will become clear in § 4, the disjunction between explicit and implicit measures is best explained by the account of trust I will suggest as opposed to the current colloquial understanding of it. To elicit group membership, participants were either shown a picture of the alleged counterpart (Stanley *et al.* 2011) – either an unknown black or white individual –, or they were given information about their counterpart's personal (picture and name) and social identity (University affiliation) (Tanis, Postmes 2005). Stanley and colleagues found that «Individuals whose IAT[1] scores reflected a stronger pro-white implicit bias were likely to offer more money to white partners than black partners, and vice versa» (Stanley *et al.* 2011: 7713). Hence, this evidence corroborates the thesis according to which implicit attitudes elicited by racial cues modulate the extent of trust granted to individuals. Analogously, Tanis and Postmes conclude that «There was less trusting behaviour when the counterpart was not personally identified and a member of the outgroup» (Tanis, Postmes 2005: 419). DeBruine's (2002) version of the trust game is slightly different from the standard one, but reaches similar conclusions. In particular, DeBruine's subjects had to decide whether to divide equally a small amount of money with another participant or to trust the other participant to divide a larger sum. As in all versions of the trust game, the other participant could also choose not to divide it. Apart from this slight difference with respect to the game used, the interesting aspect of this research is that data on actual decisions were evaluated against data of facial resemblance. To create cues of kinship, pictures of participants were manipulated

[1] The Implicit Association Test (IAT) measures the strength and speed of the association between a concept and an attribute (Greenwald *et al.* 1998). Subjects see on the top of the screen two categories, one on the right and one on the left (e.g. Black and White, Male and Female), and in the middle of the screen the word of the negative or positive feature to be attributed to one of these categories (e.g. pleasant, unpleasant, good, bad, vicious, virtuous). The amount of time spent to do the association is measured and it provides an implicit measure of participants' preferences.

using digital morphing techniques, so that the pictures that were later shown to them – as pictures of their counterparts – had different degrees of similarity with themselves. DeBruine's results are in line with both the data on in-group favouritism just reviewed and with the evolutionary data mentioned earlier. Hence, not surprisingly, facial resemblance – being a cue of kinship – enhances trust and makes subjects more inclined to trust «opponents who resembled themselves significantly more than they trusted other opponents, but did not reward trusting moves by their opponents differentially» (DeBruine 2002: 1311). This last piece of evidence will be of interest in § 4 as a further element that can be more easily explained by a two-level characterization of trust, as opposed to an ordinary one.

Interestingly, children show a similar pattern of preferences towards in-group members over out-group ones from the age of five or six years:

> Many studies have shown that preschool children prefer members of their language (Kinzler, Dupoux, Spelke 2007; Kinzler, Shutts, Dejesus, Spelke 2009), gender (Martin, Fabes, Evans, Wyman 1999; Shutts, Kinzler, McKee, & Spelke, 2009), and (to some extent) racial in-groups over out-groups (Kinzler, Spelke 2011; Kinzler *et al.* 2009) (Plötner *et al.* 2015: 162).

These data show that, as happens in adults, children also display different behaviours and attitudes when they interact with people who belong to their same group, as opposed to what they do when they interact with members of other groups.

The aim of this section was to provide some insight into a research field that has been providing evidence for quite some time now on how adults and even children have a preference for trusting and cooperating with members of their own group over members of other groups. What this literature cannot tell us, however, is whether this preference depends on the fact that subjects are more familiar with their in-groups rather than with their out-groups – as claimed by Ziv and Banaji (2012) to account for children's preferences – or on something else. To solve this problem one should resort to another line of research: that of the minimal group paradigm (§ 3).

3. *The Minimal Group Paradigm*

According to the minimal group paradigm, the preference for one's own in-group holds even when the salient groups are created arbitrarily in the lab (hence, the definition of these groups as "minimal"), as well as when subjects are given little or no time for real face-to-face interaction or when they are provided with very few cues of such a shared belonging (Plötner *et al.* 2016; Locksley *et al.* 1980; Brewer 1979; Brewer, Silver 1978; Tajfel 1974; Tajfel *et al.* 1971; Tajfel 1970). Therefore, with this line of research one can get rid of the objection according to which our preferences for in-groups may depend on familiarity. When groups are created in the lab and are not based on any visible cue, subjects are equally familiar with in-groups and with out-groups. Therefore, should the effect hold, we would need to find a reason for it without resorting to familiarity.

Plötner and colleagues (2015) have shown that 5-year-olds display a preference for members of their own minimal-group (same colour t-shirt) on multiple dimensions and even after a brief interaction. In particular, as far as trust is concerned, after children saw two puppets – one with the same colour t-shirt (in-group) and one with a different colour t-shirt (out-group) – select different boxes containing toys, they were asked to choose a box, without having the possibility to previously look into the two alternative boxes for themselves. At the age of 5, children tend to trust significantly their counterpart's choice after having cooperated with them and they show a trend to trust them more in the minimal group paradigm.

Tajfel, one of the pioneers of the minimal group paradigm, conducted several studies using this methodology (Tajfel 1974; Tajfel *et al.* 1971; Tajfel 1970). He tested 14- and 15-year-old boys with whom he pretended to divide them according to a specific criterion (i.e. he pretended to divide them among "over-estimators" and "under-estimators" of the dots that appeared on a screen, based on whether they performed "better" or "worse" at estimating the number of dots, or based on the alleged detection of a preference for Kandinsky or Klee), while they were actually divided randomly. After the division phase was completed, participants had to attribute penalties and rewards to an unknown partner – since participants were significantly older than those in Plötner and colleagues' experiment, there was no face-to-face interaction. The only thing that they were told was whether the other boy was from their own group or from the other one. Since the boys all knew each other being schoolmates, this was a tool to avoid previous friendships or hostilities to get in as confounders and to avoid any personal cue to enter the experiment. The only group identity that had to be elicited was the one Tajfel had previously given them by dividing them in two groups. As a result, participants were more kin to give more rewards and less penalties to members of their (arbitrary) group compared to the penalties and rewards they gave to members of the (arbitrary) out-group. This evidence suggests that the implicit attitudes and the biases associated with identifying with and belonging to a certain group can be triggered easily and can be elicited also by randomly selected and arbitrary groups. And this testifies to the claim that group affiliation and social categorizing are malleable. In fact, we activate the same preferences we have for long-term and stable groups (based on ethnicity, gender and the like) also for new and rather insignificant ones (e.g. wearing the same colour t-shirt as in Plötner *et al.* 2015, or supposedly preferring Kandinsky over Klee as in Tajfel 1970).

If one is worried about the possibility that in-group favouritism and the biases associated with out-groups may lead to dehumanizing members of the latter (Varga 2017), then these data constitute at the same time good and bad news, since one could arbitrarily modify who belongs to what group. The negative aspect is clearly that in-group favouritism can and is triggered without us knowing about it, and even when there are no distinctive cues supporting it. And yet, on the positive side, this malleability of in-group favouritism can also be seen as an opportunity, rather than only as a limit. Since it is so easily triggered even by simply dividing subjects in the lab, it seems at least theoretically possible to manipulate the sense of belonging so as to include people that were previously conceived of as belonging to the out-group. The upshot is that: if one aims at enlarging the scope of people whom we trust, in-group favouritism can be used as a tool to obtain such an outcome in line with the Contact Hypothesis, according to which inter-group contact would reduce stereotyping, prejudice, and discrimination (Dovidio *et al.* 2003; Allport 1954).

4. *A Two-Level Concept of Trust*

The data reviewed in the previous sections (§ 2 and 3) show that humans naturally tend to favour in-group members over out-group ones when it comes to cooperating with and trusting them. The evolutionary reasons for this are clear: as we share the goal of preserving our kins (see on this the debate on the selfish gene; Dawkins 1976; Williams 1966), we expect in-group members to act aiming at this same goal; whereas we do not expect out-groups to share it and to act in favour of it. Hence, we believe in-group members to be more trustworthy than out-group ones. While there are evolutionary reasons for this differential attitude to be in place, one could wonder whether it is also *morally justifiable* to have it. I am not convinced that it is the case that recognizing these preferential attitudes serves as a normative justification of their existence (cf. also Singer 2009: 61). Being aware of the existence of several implicit attitudes does not imply *per se* that we do not have a moral obligation to overcome them (de Lazari-Radek, Singer 2014).

On the contrary, recognizing that certain implicit drives can modulate our attitudes should be reason enough to focus on their influence and to try and limit it. In particular, one should wonder what the actual object of such influence is. If one takes trust as a unitary concept, then one is deemed to consider it subject to these drives in all its occurrences and forms. And that would mean that anytime we trust someone or something we are actually doing it *because of* these unconscious drives and not of appropriate reasons. Were it the case, trust would be deprived of any relevance in ethics since it would only be the manifestation of unconscious processes of which the subject is unaware. Just like a nervous tic, we would be unable to attribute moral responsibility to it. Trust would, thus, turn out to be amoral in all of its forms and occurrences. Hence, the data reviewed above would not be conceived of, as they should, as peculiar *amoral* cases of trusting attitudes being influenced by unconscious drives, but would be paradigmatic cases of what happens each and every time we trust someone even in situations that are actually morally relevant. On the contrary, recognizing that trust may have at least two different levels would improve our comprehension of the concept itself and would avoid considering it at the mercy of in-group favouritism at any time. It is for these reasons that I take a two-level conceptualization of trust to be more useful and to be able to preserve trust's moral dimension that would otherwise be lost because of the kind of data I have discussed. In what follows I will briefly describe what I take to be these two levels.

The first level is characterized by low-level, automatic, unconscious, and often even amoral trusting attitudes (like the ones reviewed above). It is at this level that social identities can play a role in modulating our responses. The second level, on the contrary, is the one that refers to conscious deliberations to trust someone. This is more cognitive, conscious, and deliberated. While the former is fast and refers to attitudes we often are unaware of – I might have no idea that the reason why I am more prone to expect reciprocity in a trust game from a certain counterpart (an in-group) rather than from another (an out-group) is that I have an implicit preference towards members of my own group –, the latter is the one we are interested in when attributing moral responsibility and when morally evaluating the character or behaviour of an individual. If the subject was to avow and self-attribute the reason guiding her to choose an in-group member over an out-group one, then there would be grounds to morally evaluate that attitude and the actions deriving from it. That is to say that, from the standpoint of normative ethics, one could not judge automatic trust as praiseworthy or blameworthy insofar as the individual did not choose to have such an attitude and may also be unaware of it[2]. If I tend to favour women in a trust game and I am not aware of this in-group's influence, then I should not be morally judged for having such an implicit attitude. On the contrary, if one endorses and avows her own implicit attitudes, then that person can and should be evaluate morally. For instance, if, besides unintentionally behaving in a certain way towards an ethnic out-group in a trust game in the lab, I also claim that it is right to do so, and I indulge and endorse discrimination, then I am consciously and deliberately trusting some individuals more than others based on aspects that have nothing to do with someone's trustworthiness. Skin colour or gender have, in fact, nothing to do with people's trustworthiness.

This two-level account of trust is relevant for at least three reasons. First, it allows us to explain some of the empirical evidence discussed above. For instance, by claiming that one thing are our automatic and unconscious trusting attitudes and another our deliberate ones, one is capable of accounting for the fact that, in DeBruine's experiments, subjects trusted more those who physically resembled them but did not reward trusting moves differently (2002: 1311); this distinction can also account for the absence of differences in participants' explicit assessment of how much they trust others or of whether they thought the counterpart was actually trustworthy (Stanley *et al.* 2011: 7712). Both these data can be interpreted as showing that, while the effect of implicit attitudes works perfectly well and in a very direct way as far as the automatic mechanism is concerned, it does not go through when a certain amount of reasoning and deliberate behaviour is required. When subjects have to reward others or have to assess explicitly the situation, the effect of the unconscious preference for one's own in-group members decreases or disappears. Second, this account grants that normative moral theory can be concerned with the concept of trust in its deliberate form. Deliberate trust can be morally judged

[2] This clearly depends on the notion of moral responsibility at stake and on the extent of control and awareness subjects have on their implicit attitudes and on the behaviours based on them. I have dealt with this issue in another paper, see Songhorian (2018).

– if I deliberately trust only members of my own group even if I could do otherwise, then I am and should be subjected to moral judgment – and it can be cultivated in a positive and virtuous way. Deliberately trusting others is often the morally good thing to do. Interestingly enough, while deliberation has been granted a crucial role in ethical reasoning since Aristotle (*Nicomachean Ethics*, III.3.1112a-113a), little has been said concerning its connection to the concept of trust. Distinguishing between low-level and conscious trust can, thus, also help relating these two concepts to one another. Third, this account helps getting a better grasp of the notion of trust and avoiding exaggerating or underestimating the influence implicit drives can have on it, either believing that our notion of trust has to be abandoned altogether because of the effect of implicit attitudes on *some* of its occurrences or that there is no need to modify the notion itself.

This account can also serve better than the ordinary one the purpose of understanding how the minimal group paradigm can be used to increase inter-group contact and to avoid dehumanization. Social categorizing has a direct impact only on automatic and unconscious trust, as the fastness and implicit nature of the decision to be made – in a trust game for instance – stirs our ancestors' (evolutionary) reasons to unconsciously find kin, in particular, and in-groups, in general, to be more trustworthy. Unfortunately, however, social categorizing can also have an indirect effect on conscious trust. Automatically trusting more in-group members, humans tend to acquire more and more information about previous interactions with them, rather than with individuals belonging to the out-group, thus increasing the likeliness of believing the former to be more trustworthy. Discrimination can, hence, come as a conscious and deliberate endorsement of one's experience as if experience could play the role of justifying it. Rather than recognizing that the sample of people with whom one has had interactions is limited and non-representative, some may take it as good evidence in favour precisely of the option to trust those (and only those) people more. While the effect of implicit attitudes on automatic trust is necessary, their effect on deliberate and conscious trust, happily, is not and one could also realize by reasoning that there are no good reasons to prefer in-group members over out-group ones. And yet, it is at this level that groups' malleability can be of help. If subjects are unconsciously driven to experience a sense of belonging to a group – with the implicit attitudes and preferences associated to it – within arbitrary groups composed of individuals previously conceived of as out-group members, then they will have a larger sample of different people whom they have trusted automatically. From that enlarged set of previous interactions subjects would be less likely to endorse, avow, and self-attribute explicit forms of discrimination. If one has experienced that others' trustworthiness has nothing to do with their gender or skin colour, for instance, she would less likely choose to deliberately trust only members of a certain gender or of a certain ethnic group. Group malleability can, thus, be used to directly impact automatic trust – just as much as social categorizing already does – and to indirectly impact deliberate trust – by modifying the set of previous experiences a subject has to make her inferences and to reason on in order to decide whom to trust. By showing how malleable groups are, the minimal group paradigm obtains humanization, which is the exact contrary of the dehumanization that leads to stereotyping, prejudice, and discrimination. The minimal group paradigm can be, therefore, used as a practical tool to make people from different groups enter in contact with each other (in line with the Contact Hypothesis).

5. *Conclusion*

The aim of this paper was to provide some evidence in favour of the need for a revision of our ordinary concept of trust. Evidence from studies on the investment game – also known as the trust game – and on strangers' allocation of money to an out-group or an in-group (Platow *et al.* 2012; Stanley *et al.* 2011; Foddy *et al.* 2009; Güth *et al.* 2008; Tanis, Postmes 2005; DeBruine 2002) suggest that our group identity modulates the extent to which we trust others – favouring member of our own group over members of other groups (§ 2). To get rid of the objection according to which the data collected could depend on the fact that subjects are more familiar with in-groups than with out-groups, I have resorted to the minimal group paradigm (§ 3). The evidence from this important line of research does not only eliminate the possibility of claiming that familiarity plays such a role in shaping humans' attitudes and behaviours towards in-groups and out-groups, but it also pushes the results further by claiming that the effect is not evident only in cases in which the identities at stake are strong and entrenched ones (like ethnicity or gender, for instance), but also when groups are created arbitrarily in the lab (Plötner *et al.* 2015; Locksley *et al.* 1980; Brewer 1979; Brewer, Silver 1978; Tajfel 1974; Tajfel *et al.* 1971; Tajfel 1970). Furthermore, the data from the minimal group paradigm are particularly revealing of the malleability and of the almost immediate impact social categorizing can have on automatic trusting attitudes. While this clearly bears huge risks of malevolent manipulation, this malleability can also be an opportunity: it seems at least theoretically possible to manipulate the sense of belonging – and the automatic trust that follows from it – so as to include people that were previously conceived of as belonging to other groups. The upshot is the following: if one aims at enlarging the scope of people whom we trust to achieve humanization as opposed to dehumanization, our own biases can be used as a practical tool to obtain such an outcome.

These two lines of research have been used to show that there are several implicit drives that can modulate our trusting attitudes even if we do not know about them. Recognizing this leads to a revision of our ordinary conceptualization of trust – that I briefly discussed in § 1. Without such a revision, in fact, we would run two symmetrical risks: either exaggerating or underestimating the influence of implicit drives. If we exaggerate their impact and consider them applicable to a unitary concept of trust, then it seems that the notion of trust can have no role in ethics as there is nothing deliberate about it. On the contrary, if one aims at maintaining exactly the concept of trust we currently use – with its indistinction between amoral and moral applications –, then one needs to refute the possibility of implicit attitudes having any impact on trust altogether. To avoid these symmetrical risks, I proposed a two-level characterization of trust that would better serve the purposes of accounting for the data here discussed and for the role trust can and should play in ethics (§ 4). I have argued that the kind of trust that is subject to modulation and distortion by these unconscious drives is not the same kind we are interested in when we do moral philosophy. That is, one has to distinguish between low-level and automatic trust – the amoral one that can easily and unconsciously be biased – and more cognitive, conscious and deliberated forms of trust – those that are actually morally relevant. Since the latter is an attitude that the agent reflectively endorses and self-avows, it manifests one's moral personality and the agent can be deemed fully responsible for the actions stemming from deliberate trust. On the contrary, these features do not apply in the case of automatic trust. Hence one can conclude that, even though a moral theory would have to be primarily concerned with deliberate and conscious trust, there is still room for using our limitations to our benefit by unconsciously modulating automatic trust since it would modify the set of experiences an individual has as a basis for future inferences and expectations. This would not lead *per se* to huge differences in moral behaviour, but it might spread a positive bias that a moral agent would try to pick up and cultivate at a more conscious level.

In conclusion, in this paper I have focused in particular on the impact implicit drives have on interpersonal trust leaving aside other relevant issues that deserve more attention that I could devote to them here. For instance, I have not delved into how unconscious drives can impact other forms of trust – as trust for institutions or self-trust. Furthermore, I have not focused on whether automatic forms of trust for the dearest and nearest have *any* moral relevance or consequence *per se*. Is it a good character trait to automatically trust others? Does it lead to some interpersonal virtue? Or, on the contrary, being unconsciously trusting is completely out of the domain of moral action? While I have claimed that deliberate trust is the properly moral one, this does not imply that having a trusting character cannot have some moral consequences (as, for instance, leading more easily to some virtues like benevolence). These issues would be the object of further analysis and research.

References

Allport G.W. (1954), *The Nature of Prejudice*, Addison-Wesley, New York.

Appiah K.A. (2008), *Experiments in Ethics*, Harvard University Press, Cambridge.

Aristotle (2000), *Aristotle: Nicomachean Ethics*, R. Crisp (ed.), Cambridge University Press, Cambridge.

Baier A.C. (1986), *Trust and Antitrust*, in «Ethics», vol. 96, pp. 231-260.

Berg J., Dickhaut J., McCabe K. (1995), *Trust, Reciprocity, and Social History*, in «Games and Economic Behavior», vol. 10, pp. 122-142.

Bernhard H., Fischbacher U., Fehr E. (2006), *Parochial Altruism in Humans*, in «Nature», vol. 442, n. 7105, pp. 912-915.

Brewer M.B. (1979), *In-group Bias in the Minimal Intergroup Situation: A Cognitive-Motivational Analysis*, in «Psychological Bulletin», vol. 86, pp. 307-324.

Brewer M.B., Silver M. (1978), *Ingroup Bias as a Function of Task Characteristics*, in «European Journal of Social Psychology», vol. 8, pp. 393-400.

Dasgupta N., Banaji M.R., Abelson R.P. (1999), *Group Entitativity and Group Perception: Associations between Physical Features and Psychological Judgment*, in «Journal of Personality and Social Psychology», vol. 77, n. 5, pp. 991-1003.

Dawkins R. (1976), *The Selfish Gene*, Oxford University Press, Oxford.

de Lazari-Radek K., Singer P. (2014), *The Point of View of the Universe. Sidgwick and Contemporary Ethics*, Oxford University Press, Oxford.

DeBruine L.M. (2002), *Facial Resemblance Enhances Trust*, in «Proceeding of the Royal Society of London B», vol. 269, pp. 1307-1312.

Doris J. (1998), *Persons, Situations, and Virtue Ethics*, in «Noûs», vol. 32, pp. 504-530.

Doris J. (2002), *Lack of Character: Personality and Moral Behavior*, Cambridge University Press, Cambridge.

Doris J. (2010), *Heated Agreement: Lack of Character as Being for the Good*, in «Philosophical Studies», vol. 148, pp. 135-146.

Dovidio J.F., Gaertner S.L., Kawakami K. (2003), *Intergroup Contact: The Past, Present, and the Future*, in «Group Processes & Intergroup Relations», vol. 6, n. 1, pp. 5-21.

Foddy M., Platow M.J., Yamagishi T. (2009), *Group-Based Trust in Strangers: The Role of Stereotypes and Expectations*, in «Psychological Science», vol. 20, n. 4, pp. 419-422.

Greene J. (2013), *Moral Tribes: Emotion, Reason, and the Gap between Us and Them*, Atlantic Books, London.

Greenwald A.G., McGhee D.E., Schwartz J.L.K. (1998), *Measuring Individual Differences in Implicit Cognition: The Implicit Association Test*, in «Journal of Personality and Social Psychology», vol. 74, pp. 1464-1480.

Güth W., Levati M.V., Ploner M. (2008), *Social Identity and Trust - An Experimental Investigation*, in «Journal of Socio-Economics», vol. 37, pp. 1293-1308.

Hamilto D.L., Sherman S.J., Lickel B. (1998), *Perceiving Social Groups: The Importance of the Entitativity Continuum*, in C. Sedikides, J. Schopler, C.A. Insko (eds.), *Intergroup Cognition and Intergroup Behavior*, Erlbaum, Mahwah, pp. 47-74.

Hamilton W.D. (1964), *The Genetical Evolution of Social Behaviour II*, in «Journal of Theoretical Biology», vol. 7, pp. 17-52.

Harman G. (1999), *Moral Philosophy meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error*, in «Proceedings of the Aristotelian Society», vol. 99, pp. 315-331.

Harman G. (2000), *The Nonexistence of Character Traits*, in «Proceedings of the Aristotelian Societyin «, vol. 100, pp. 223-226.

Harman G. (2001), *Virtue Ethics without Character Traits*, in A. Byrne, R. Stalnaker, and R. Wedgwood (eds), *Fact and Value*, MIT Press, Cambridge, pp. 117-127.

Harman G. (2003), *No Character or Personality*, in «Business Ethics Quarterly», vol. 13, pp. 87-94.

Harman G. (2009), *Skepticism about Character Traits*, in «The Journal of Ethics», 13, pp. 235-242.

Hawley K. (2012), *Trust: A Very Short Introduction*, Oxford University Press, Oxford.

Herdova M. (2016), *What You Don't Know Can Hurt You: Situationism, Conscious Awareness, and Control*, in «Journal of Cognition and Neuroethics», vol. 4, n. 1, pp. 45-71.

Huang J.L. (2014), *Does Cleanliness Influence Moral Judgments? Response Effort Moderates the Effect of Cleanliness Priming on Moral Judgments*, in «Frontiers in Psychology», vol. 5, n. 1276, pp. 1-8.

Johnson N.D., Mislin A.A. (2011), *Trust Games: A Meta-Analysis*, in «Journal of Economic Psychology», vol. 32, n. 5, pp. 865-889.

Kropotkin P. (1908), *Mutual Aid: A Factor of Evolution*, William Heineman, London.

Levine M., Prosser A., Evans D., Reicher S. (2005), *Identity and Emergency Intervention: How Social Group Membership and Inclusiveness of Group Boundaries Shape Helping Behavior*, in «Personality & Social Psychology Bulletin», vol. 31, n. 4, pp. 443-453.

Levy N. (2017), *Implicit Bias and Moral Responsibility: Probing the Data*, in «Philosophy and Phenomenological Research», vol. 94, pp. 3-26.

Liljenquist K., Zhong C.-B., Galinsky A.D. (2010), *The Smell of Virtue: Clean Scents Promote Reciprocity and Charity*, «Psychological Science», vol. 21, n. 3, pp. 381-383.

Locksley A., Ortiz V., Hepburn C. (1980), *Social Categorization and Discriminatory Behavior: Extinguishing the Minimal Intergroup Discrimination Effect*, in «Journal of Personality and Social Psychology», vol. 39, pp. 773-783.

McLeod C. (2015), *Trust*, in E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Fall 2015 Edition)*, http://plato.stanford.edu/archives/fall2015/entries/trust/ [September 22nd, 2018].

Platow M.J., Foddy M., Yamagishi T., Lim L., Chow A. (2012), *Two Experimental Tests of Trust in In-group Strangers: The Moderating Role of Common Knowledge of Group Membership*, in «European Journal of Social Psychology», vol. 42, pp. 30-35.

Plötner M., Over H., Carpenter M., Tomasello M. (2015), *The Effects of Collaboration and Minimal-Group Membership on Children's Prosocial Behavior, Liking, Affiliation, and Trust*, in «Journal of Experimental Child Psychology», vol. 139, pp. 161-173.

Plötner M., Over H., Carpenter M., Tomasello M. (2016), *What Is a Group? Young Children's Perceptions of Different Types of Groups and Group Entitativity*, in «PLoS ONE», vol. 11, n. 3, pp. e0152001.

Schnall S., Benton J., Harvey S. (2008), *With a Clean Conscience: Cleanliness Reduces the Severity of Moral Judgments*, in «Psychological Science», vol. 19, n. 12, pp. 1219-1222.

Simpson T.W. (2012), *What Is Trust?*, in «Pacific Philosophical Quarterly», vol. 93, pp. 550-569.

Singer P. (2009), *The Life You Can Save. How to Play Your Part in Ending World Poverty*, Picador, London.

Songhorian S. (2018), *Implicit Attitudes' Challenge to Moral Responsibility*, in «Notizie di Politeia», vol. XXXIV, n. 131, pp. 73-88.

Stanley D.A., Sokol-Hessner P., Banaji M.R., Phelps E.A. (2011), *Implicit Race Attitudes Predict Trustworthiness Judgments and Economic Trust Decisions*, in «Proceedings of the National Academy of Sciences of the United States of America», vol. 108, n. 19, pp. 7710-7715.

Tajfel H. (1970), *Experiments in Intergroup Discrimination*, in «Scientific American», vol. 223, pp. 96-102.

Tajfel H. (1974), *Social Identity and Intergroup Behaviour*, in «Social Science Information», vol. 13, pp. 65-93.

Tajfel H., Billig M.G., Bundy R.P., Flament C. (1971), *Social Categorization and Intergroup Behaviour*, in «European Journal of Social Psychology», vol. 1, pp. 149-178.

Tanis M., Postmes T. (2005), *A Social Identity Approach to Trust: Interpersonal Perception, Group Membership and Trusting Behaviour*, in «European Journal of Social Psychology», vol. 35, pp. 413-424.

Varga S. (2017), *The Case for Mind Perception*, in «Synthese», vol. 194, pp. 787-807.

Williams G.C. (1966), *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought*, Princeton University Press, Princeton.

Wynne-Edwards V.C. (1962), *Animal Dispersion in Relation to Social Behaviour*, Oliver & Boyd, Edinburgh.

Xu X., Zuo X., Wang X., Han S. (2009), *Do You Feel My Pain? Racial Group Membership Modulates Empathic Neural Responses*, in «The Journal of Neuroscience: The Official Journal of the Society for Neuroscience», vol. 29, n. 26, pp. 8525-8529.

Yzerbyt V.Y., Rogier A., Fiske S.T. (1998), *Group Entitativity and Social Attribution: On Translating Situational Constraints into Stereotypes*, in «Personality and Social Psychology Bulletin», vol. 24, n. 10, pp. 1089-1103.

Ziv T., Banaji M.R. (2012), *Representations of Social Groups in the Early Years of Life*, in S.T. Fiske, C.N. Macrae (eds.), *The Sage Handbook of Social Cognition*, Sage, London, pp. 372-389.

Abstract

Several empirical evidences suggest that our group identity modulates our trusting attitudes, even when groups are created arbitrarily in the lab. Hence, group are malleable entities. While it clearly bears huge risks of malevolent manipulation, this malleability can also be an opportunity: it seems at least theoretically possible to manipulate the sense of belonging – and the automatic trust that follows from it – so as to include people that were previously conceived of as belonging to other groups.

I will, thus, investigate two lines of research to be used to show that there are several implicit drives that actually modulate our trusting attitudes. From this, a revision of our ordinary conceptualization of trust seems necessary. Hence, I proposed a two-level characterization of trust that would better serve the purposes of accounting for the data discussed and for the role trust can and should play in ethics.

Keywords: minimal group paradigm; trust; social categorizing; ethics.

Sarah Songhorian
Università Vita-Salute San Raffaele
songhorian.sarah@unisr.it