The key to group fitness: The presence of another synchronizes moral attitudes and neural responses during moral decision-making

Chenyi Chen, Róger Marcelo Martínez, Yawei Cheng

PII: S1053-8119(20)30219-6
DOI: https://doi.org/10.1016/j.neuroimage.2020.116732
Reference: YNIMG 116732

To appear in: NeuroImage

Received Date: 5 June 2019
Revised Date: 6 March 2020
Accepted Date: 9 March 2020

Please cite this article as: Chen, C., Martínez, Ró.Marcelo., Cheng, Y., The key to group fitness: The presence of another synchronizes moral attitudes and neural responses during moral decision-making, NeuroImage (2020), doi: https://doi.org/10.1016/j.neuroimage.2020.116732.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2020 Published by Elsevier Inc.
The key to group fitness: The presence of another synchronizes moral attitudes and neural responses during moral decision-making

Chenyi Chen\textsuperscript{1,2,3,4}, Róger Marcelo Martínez\textsuperscript{2}, Yawei Cheng\textsuperscript{1,5,6 CA}

1. Department of Physical Medicine and Rehabilitation, National Yang-Ming University Hospital, Yilan, Taiwan.

2. Graduate Institute of Injury Prevention and Control, College of Public Health, Taipei Medical University, Taipei, Taiwan.

3. Research Center of Brain and Consciousness, Shuang-Ho Hospital, Taipei Medical University, New Taipei City, Taiwan.

4. Graduate Institute of Mind, Brain and Consciousness, College of Humanities and Social Sciences, Taipei Medical University, Taipei, Taiwan.

5. Institute of Neuroscience and Brain Research Center, National Yang-Ming University, Taipei, Taiwan.

6. Department of Education and Research, Taipei City Hospital, Taipei, Taiwan.

\textbf{Corresponding author:} Prof. Yawei Cheng

Institute of Neuroscience and Brain Research Center

National Yang-Ming University
No. 155, Sec. 2, St. Linong, Dist. Beitou
Taipei 112, Taiwan, ROC.

Tel: 886-2-28267912
Fax: 886-2-28264903
Email: ywcheng2@ym.edu.tw

Running title: The confederate alters moral attitudes.

Number of words: abstract (207 words, 5 key words), text (6426), 2 tables, and 5 figures.
Abstract

Morality encompasses a set of norms that originate from the group’s consensus and cultural evolution. Accordingly, the mere presence of another person is very well capable of shaping moral judgment and decision-making in a way that benefits group coherence. Here, we measured justice sensitivity (JSI), implicit moral attitudes (mIAT), and neural activity during mental simulation of interpersonal morally-laden behaviors (helping and harming) when participants were either alone or when they thought a confederate was present. Both JSI and mIAT, as well as various brain networks, were modulated and, further, synchronized by the presence of a confederate. Participants with lower scores on the mIAT and JSI enhanced their moral attitudes when they were in the presence of a confederate. This change was driven by increased signal in the amygdala and anterior insula when the low mIAT participants mentally simulated harming another person, but was effected by decreased activity in the dorsomedial- and dorsolateral-prefrontal cortex in the high JSI participants. The presence of another significantly impacts moral attitudes as well as neural correlates underlying moral behavior. Together, the results support the view that both individual dispositions and social influence shape and synchronize people’s moral computations, and fits with the theoretical perspective that morality has evolved to promote group fitness.
Keywords: social influence; moral attitudes; moral actions; morality; justice sensitivity

1. Introduction

Morality encompasses a set of behaviors and attitudes originating from group consensus and whose function is for such groups to successfully evolve into stable societies, thus subsequent reinforcement and sanction of these norms is crucial (Ayala, 2010). Consequently, it has been observed in many animal species (including humans) that the mere presence of another conspecific strongly exerts social influence and affects most behaviors, whether they are basic (e.g., food consumption) or sophisticated (e.g., attention, memorization, categorization, etc.) (Guerin, 2009). It is also very well known that these social variables -e.g., the presence and behavior of others, but also media exposure, and even hearsay (Dasgupta, 2013) - are very well able to influence moral attitudes.

One of the most famous examples is the bystander effect, which refers to a decrease in the helping behavior of each individual within a group of people, who simultaneously witness a person in need of assistance (Darley and Latane, 1968). This
phenomenon shows how the inhibiting effect offered by the presence of others is objectively generalized, and demonstrates that attentiveness to contextual events decreases by this same presence as well. Moreover, research has shown that an increase in the number of people in a group when witnessing an emergency decreases the activity in those brain areas typically engaged during action preparation (Hortensius and de Gelder, 2014), as the belief that it is another who will end up offering assistance is shared and synchronized between individuals.

Alongside, while research on social presence comparing an ‘alone’ versus a ‘witnessed by another’ condition has been performed (Andreoni and Petrie, 2004; Jung et al., 2018; Lee et al., 2018), it comes with certain limitations. First, these studies place the participants under the passive surveillance of an experimenter while performing a behavioral task during the ‘witnessed by another’ condition, whereas the present study attempts for the confederate to actively perform the task at the same time with the participant. Second, the tasks mainly consisted in indirect helping through economic decision-making (buying goods whose proceeds would result in financial aid for people in need, or donating money), while this study deals not only with direct helping behaviors (e.g., assisting someone with carrying a bag, aiding someone by dressing their wounds, etc.), but also with direct harming.
behaviors. Third, none of the studies took into consideration implicit social attitudes. The latter of which are automatic attitudes, able to bypass the attentive vigilance of consciousness (thus resistant to free will in nature), and mistakenly thought to be immutable. In fact, implicit attitudes have been observed to be surprisingly malleable, and are modulated by the communities and environments in which the individual is ingrained at the moment. Whereas explicit attitudes are consciously accessible, controlled and, therefore, changeable depending on the context (Dasgupta, 2013). Contrary to implicit attitudes, it is because of the sense of ownership that explicit attitudes themselves confer that they are susceptible to change, as individuals become aware of their own actions being possibly surveilled (Ernest-Jones et al., 2011).

Furthermore, previous research suggests that moral attitudes play a role in the regulation of individual behavior within social systems (Ellemers et al., 2013). For example, and taking into account the bystander effect, whether it happens due to such proposed mechanisms as pluralistic ignorance, diffusion of responsibility, or evaluation apprehension (for a meta-analysis, please see Fischer et al., 2011), all of these reflect interpretations that appeal to high level cognitive notions and provide a top-down explanation for social behavior. While these interpretations undoubtedly
shed some light regarding group influences on individual behaviors, the broader issue is how various influences affect the perception and interpretation of stimuli that have the potential to evoke up-regulation.

A substantial body of neuroimaging research suggests that moral cognition arises from the integration of both rationally-based (cognitive) and emotionally-based (affective) subsystems (Greene et al., 2004; Gubbins and Byrne, 2014; Kahane, 2015). The cognitive reasoning process is associated with activity in the dorsolateral prefrontal cortex (dLPFC) (Buckholtz et al., 2008), dorsomedial prefrontal cortex (dmPFC) (Forbes and Grafman, 2010), and posterior superior temporal sulcus/temporoparietal junction (pSTS/TPJ) (Young and Dungan, 2012). The automatic affective process engages the amygdala (Pascual et al., 2013), anterior insula (aIN) (Schaich Borg et al., 2008), and ventromedial prefrontal cortex (vmPFC) (Koenigs et al., 2007).

While dispositional (explicit) justice sensitivity (JSI) is associated with moral evaluations as well as with dmPFC activation (Yoder and Decety, 2014a), implicit moral attitudes, as indicated by the morally-laden implicit association test (mIAT), are able to predict actor-observer asymmetry in blame attribution and orbitofrontal
cortex activation (Chen et al., 2020). However, how the presence of a confederate can alter one’s dispositional (explicit) and implicit moral attitudes remains elusive. Here, we show how the presence of another modulates the interplay between the dynamic neural subsystems underpinning moral cognition and behavior, as well as its effect on individual differences concerning implicit moral attitudes and explicit justice sensitivity. In order to provide a more comprehensive understanding of the complex neural dynamics involved in morality, we investigated whether the presence of a confederate, who was doing the same moral tasks alongside the participants, could alter the participants’ moral attitudes and behaviors. Since morality as a set of norms to be sanctioned originates from the group’s consensus and cultural evolution (Ayala, 2010), we hypothesized that the presence of a confederate would modify moral cognition by means of the synchronization of implicit moral attitude- and explicit justice sensitivity-related discrepancies between individuals, and that this synchronization would further be able to be confirmed at a neural level. Specifically, the participant groups’ (as divided by their differing mIAT and JSI scores) attitudes would become comparable as an effect of the presence of a confederate.

2. Materials and Methods

2.1. Participants
Fifty healthy (25 females) Han Chinese volunteers with their ages spanning from 20 to 36 years old were recruited from the college campus and enrolled in the study after providing written informed consent. After which they received a monetary compensation of five-hundred new Taiwan dollars. Participants were enrolled as a result of advertising efforts in a local community through a recruitment coordinator. Data collection was conducted from January-February and July-August 2017 (i.e. the summer and winter breaks along the Taiwanese academic calendar). In order to maximize the ecological validity and ensure the effects of the experimental manipulation — i.e., the belief in the presence of a confederate who is under the same conditions as the participants —, as well as minimize the inter-individual variation in social desirability, we only enrolled the participants who showed no doubts in regards to the confederate’s true identity (namely, those who believed the confederate to be a real participant like him/her rather than a fake actor or another experimenter). As a result of doubts raised in regards to the confederate during the debriefing stage, as well as due to excessive head movement during MRI scanning, 10 subjects (4 females) were excluded from the data analyses [final sample: N = 40 (21 females), mean age ± SD: 23.5 ± 2.2]. All participants had normal or corrected-normal visual acuity. None of them had any history of neurological or psychiatric disorders, and all were free of medication at the time of testing. The study was approved by the
Ethics Committee of National Yang-Ming University.

2.2. Visual stimuli

45 validated animations, assessed by previous fMRI studies in which participants were asked to rate the outcome of the actions using a scale going from “no pain” to “extremely painful” and from “no effect” to “extreme unpleasantness,” as target words on the Facial Pain Scale-Revised (FPS-R) (Bieri et al., 1990), were presented to participants as the stimuli for the morally-laden scenario task (Akitsuki and Decety, 2009; Decety et al., 2008; Decety and Porges, 2011). Each animation was comprised of three images with a duration of 1000, 200, and 1000 milliseconds, respectively, and portrayed the following scenarios: (1) a person who was performing an action that physically harmed another person (harming); and (2) a person who was alleviating physical pain from a suffering person (helping). One additional baseline stimuli set depicted people carrying out an action that was irrelevant to another person (neutral). The faces of the protagonists were not visible as to ensure there was no emotional reaction from the participants. In order to help participants successfully take on the perspective of the active role, a handheld button needed to be pressed once before the action proceeded. That is, the subject would see the first
image of the action, and proceed to press the button in order to induce the remaining two images to play out.

2.3. Procedures

Participants’ implicit moral attitudes (mIAT) and Justice sensitivity (JSI) were assessed before fMRI scanning. During fMRI scanning, in order to help the participants successfully take on the perspective of the active role on actions that carry different moral consequences (being them helping, harming, or neutral in nature), participants engaged in the button-trigger task, where a handheld button needed to be pressed once for the action to proceed. The participant would observe the first image of the action clip (with no time limit set for this 1st image), then would have to press the button to induce the remaining two images to play out. Participants were told to press the button in a self-paced manner, without them being encouraged to press it as soon as possible (mean RT = 1228 ± 732 ms). Each participant underwent each condition (helping, harming and neutral), with the order of these blocks being randomized but not counterbalanced across participants in a session where they were alone, and another session where they had the confederate. The two sessions were counterbalanced and separated with an interval of at least one week apart.
In the alone session, participants performed the mIAT, filled in the JSI, underwent fMRI scanning, and completed moral evaluations all by themselves in the presence of the experimenter. In the confederate session, participants were first introduced to the confederate. Then, both of them were simultaneously given the instructions by the experimenter in the same room. They then performed the mIAT and filled in the JSI together in the same room side-by-side, two meters apart from each other, so that their performance could not be seen by each other. The confederate was the same female adult (aged 33) across all participants. During fMRI scanning, a colored dot was shown during the presentation of morally-laden scenarios to indicate the exact time when the confederate allegedly pressed the button, and convincing the participants that they were doing the moral behavior task together with someone else outside the scanner via a connected computer.

After fMRI scanning, participants underwent the same procedures with the same situations (alone vs. confederate) that they did in the scanner. After each trial, they were asked to evaluate how much blame or praise they deserved themselves for “enacting” these morally-laden behaviors by using a computer-based 7-point visual analogue scale. Thereafter, we gave a short debriefing to the participants.
2.3.1. Implicit moral attitudes (mIAT)

Both word and visual animations were used for creating the stimuli. The word stimuli included 26 extremely pleasant and 26 extremely unpleasant words, selected from highly frequent used Chinese words (Chen et al., 2002). The animation stimuli consisted of 47 clips depicting everyday dyadic interactions, in which an action that is carried out directly toward another person is either characterized as a moral or immoral action (Yoder and Decety, 2014a, b). The mIAT followed the experimental design proposed by Greenwald and his colleagues (1998). The procedures involved a series of five discrimination blocks. The accuracy rate and reaction time (RT) during Block 3 and 5 were recorded. The five blocks were:

Block 1 started with the "initial target-concept discrimination". Participants categorized the clips as moral (right response key) or immoral (left response key).

Block 2 was termed as "attribute discrimination". Participants categorized words as negative (right response key) or positive (left response key).

Block 3 "target-concept discrimination" combined Block 1 and 2 with clips and words randomly presented in alternative trials. The moral clips share a right response
key with negative words, and immoral clips shared a left response key with positive words (moral-negative/immoral-positive).

Block 4 was known as "reversed target-concept discrimination". Participants learned a reversed response assignment for block 1 and judged if clips were moral (left response key) or immoral (right response key).

Block 5 combined Block 2 and 4. The immoral clips shared a right response key with negative words, and moral clips shared a left response key with positive words (immoral-negative/moral-positive).

Two block sequences (12345/42513) were counterbalanced to control the sequential effect. Half of the participants experienced the incongruent block first, and the other half completed the congruence block first. Blocks 1, 2, and 4 contained 20 trials each, whereas Blocks 3 and 5 (incongruent and congruent blocks) contained 40 trials each.

Words and/or clips within each block were presented in a random order. Each trial consisted of a presentation of a 1000-ms fixation followed by a stimulus. The stimuli were viewed on a computer and administered using E-Prime version 2.0 software (Psychology Software Tools). Participants were instructed to classify each word or clip as fast and accurately as possible. There were ten practice trials before participants started Block 3 or Block 5.
The mIAT performance, as indexed by D scores, represents the implicit moral attitudes. The D score was calculated by subtracting the mean RT of congruent (immoral-negative) blocks from that of incongruent (immoral-positive) blocks and divided it by the pooled standard deviation across the two blocks (Greenwald et al., 1998; Nosek et al., 2014). The higher D scores might come from higher RT for the incongruent blocks or lower RT for the congruent blocks.

Because the implicit association test relies on the RT differentials, which are highly sensitive to outliers and extreme values, reaction times exceeding two times the standard deviation from the subject means were excluded from the set of valid responses (outliers accounted for less than 3% of all the responses). Additionally, extreme responses – either very slow or very fast – can indicate inattention to the task performance rules. Exclusion criteria were applied for RT faster than 300-ms and slower than 6000-ms cut-off boundary response latency (extreme responses accounted for less than 1% of all responses) (Nosek et al., 2014).

2.3.2. Dispositional justice sensitivity (JSI)

The JSI is a self-reported psychometric measure that assesses justice sensitivity from four different perspectives: as a victim, e.g., it bothers me when others receive
something that ought to be mine; as an observer, e.g., it bothers me when someone
gets something they don’t deserve; as a beneficiary, e.g., it disturbs me when I
receive what others ought to have; and as a perpetrator, e.g., it gets me down when
I take something from someone else that I don’t deserve. Each perspective has 10
items, with each item rated on a seven-point scale from 0 (not at all) to 7 (strongly
agree), and which indicate an individual’s disposition to react to unfair situations. In
total, with all perspectives’ scores combined, a global score between 0 to 280 is
obtained. This total score indexes the individual’s perceptual threshold of norm
violation and injustice (Schmitt et al., 2010).

2.4. fMRI scanning

Participants underwent two fMRI scanning sessions (alone, and with the presence of
a confederate) on separate days, at least one week apart. Stimuli were presented
with the E-prime software (Psychology Software Tools, Inc., Pittsburgh, PA) and an
fMRI compatible goggle (VisualStim Controller, Resonance Technology Inc.) in a 2 x 3
factorial design: session (alone vs. confederate) x scenario (harming vs. helping vs.
neutral).

The scanning followed a block design (23.1 ± 0.7 s ON/ 13.2 ± 4.4 s OFF). Each
session had two runs. Each run consisted of 6 ON blocks (2 harming, 2 helping, and 2 neutral scenarios) intermixed with 6 OFF blocks. Each ON block consisted of five trials (duration $2428 \pm 732$ ms each) and five inter-stimulus intervals (duration 2200-ms each) with a fixation cross presented against a gray background. The sequence of the blocks (harming, helping, neutral) was randomized within each run. The session order (alone vs. confederate) was counterbalanced across subjects.

Scanning was performed on a 3T Siemens Magnetom Trio-Tim magnet. For functional changes, changes in blood oxygenation level-dependent (BOLD) T2* weighted MR signal were collected along the AC–PC plane using a gradient echo-planar imaging (EPI) sequence (TR = 2200 ms, TE = 30 ms, FOV = 220 mm, flip angle = 90°, matrix = 64 × 64, 36 transversal slices, voxel size = $3.4 \times 3.4 \times 3.0$ mm$^3$, no gap). High-resolution structural T1-weighted images were acquired using a 3D magnetization-prepared rapid gradient echo sequence (TR = 2530 ms, TE = 3.5 ms, FOV = 256 mm, flip angle = 7°, slice thickness = 1 mm, matrix = 256 × 256, no gap).

### 2.4.1. fMRI data analysis

Functional MRI data was processed with SPM12 (Wellcome Department of Imaging Neuroscience, London, UK) in MATLAB 9.0 (MathWorks Inc., Sherborn, MA, USA).
Structural T1 images were coregistered to the mean functional images, and a skull-stripped image was created from the segmented gray matter, white matter, and CSF images. These segmented images were combined to create a subject-specific brain template. EPI images were realigned and filtered (128 s cutoff), then coregistered to these brain templates, normalized to MNI space, and smoothed (8 mm FWHM). All participants who completed scanning had less than 1 voxels of in-plane motion. A two-level approach for mixed-design fMRI data was adopted. A voxel-by-voxel multiple regression analysis of expected signal changes for each of the 3 block categories (harming vs. helping vs. neutral), which were constructed using the hemodynamic response function, was applied to the preprocessed images for each participant. Individual subject data were analyzed using a fixed-effects model. Boxcar regressors, which represented the occurrence of each of the 3 block categories, were used to model condition effects at the subject level. At the first level analysis, the three conditions (harming, helping, and neutral) were modeled separately with a duration of the length of each ON block beginning at the onset of the first image of the action. The null event (fixation) was modeled with the duration $13.2 \pm 4.4$ s. Linear contrasts were applied to obtain parameter estimates. Movement parameters from the realignment output were included as regressors of no interest. The resulting first-level contrast images were then entered into an analysis of
variance (ANOVA): 2 (group: High vs. Low D) x 2 (session: alone vs. confederate) x 2 (action: harming vs. helping). Whole brain activations were corrected for multiple comparisons family-wise error (FWE) rate at $P < 0.05$.

Hemodynamic responses in specific regions of interest (ROIs) were analyzed using the MarsBaR toolbox (http://marsbar.sourceforge.net/) implemented in SPM12. The ROIs for the amygdala (x -22, y -2, z -24) and aIN (-30, 20, 4) were defined as a 4-mm spherical region centered on the coordinates determined on the basis of neuroanatomical atlases as well as meta-analyses of morality and empathy (Bzdok et al., 2012; Lamm et al., 2011), whereas those of the vmPFC (10, 42, -18), pSTS/TPJ (56, -50, 18), dlPFC (42, 30, 26), and dmPFC (0, 54, 36) were defined as a 4-mm spherical region based on prior fMRI studies of moral evaluation with similar stimuli (Decety et al., 2012; Decety and Porges, 2011; Yoder and Decety, 2014a). The individual mean parameter estimates (beta values) were subject to an ANOVA to test main effects of group, session, and scenario, as well as group-by-session, session-by-action, group-by-action, and group-by-session-by-action interactions. For statistical analyses of the ROI data, SPSS was used. Statistical power (1-β) was estimated by G*Power 3.1 software (Faul et al., 2007).
3. RESULTS

Forty healthy (21 females) participants with their ages spanning from 20 to 36 years of age (mean ± SD: 23.5 ± 2.2) were included for analysis. The D scores from the mIAT ranged from 0.07 to 1.16 (mean ± SD: 0.6 ± 0.23) in the alone session and from 0.21 to 1.41 (0.78 ± 0.27) in the confederate condition. The scores of dispositional justice sensitivity (JSI) ranged from 71 to 256 (175.86 ± 35.04) in the alone session and from 119 to 246 (177.35 ± 29.58) in the confederate condition (Table 1).

3.1. Implicit moral attitudes (mIAT)

The presence of a confederate significantly increased the D scores of the mIAT ($t_{39} = 3.33, P = .002, (1-\beta) = 99.73\%$). The mIAT modulation varied as a function of the individual baseline D scores where it was derived from the alone condition. $\Delta D$ scores (confederate – alone) were negatively correlated with the baseline D scores in the alone condition ($r = -0.61, P < .001$) (Fig. 1A). Accordingly, we further divided groups of relatively high and low D based on median split of the baseline D scores (Median = 0.65; Mean ± SD = 0.62 ± 0.23). We normalized the individual $\Delta D$ induced by the presence of a confederate and performed a non-parametric one-sample-t test (Wilcoxon signed rank tests) separately for these groups. Results showed that the group with lower baseline D scores (Low D: $n = 20$) had a marginally significant
increase in \( \Delta D \) (\( Z = 1.89, P = .059 \)), whereas the group with higher baseline D scores (High D: \( n = 20 \)) was not significantly modulated by the presence of a confederate (\( Z = 0.99, P = .32 \)) (Fig. 1B). Neither the gender effect nor the related interaction was significant (all \( P > 0.1 \)).

Given that the confederate presence did not cause a universal effect across participants and its modulation varied as a function of individual baseline mIAT, we consequently conducted a mixed-design ANOVA with a within-subject variable (session: alone vs. confederate) and a between-subject variable (group: High vs. Low D) for the mIAT. There was an interaction between group and session (\( F_{1, 38} = 5.57, P = .024, \eta_p^2 = 0.128, (1-\beta)^{-100}\% \)). Post hoc analyses revealed that the presence of a confederate increased D scores only for the Low D (alone vs. confederate: 0.42 ± 0.03 vs. 0.72 ± 0.06), but not for the High D group (0.79 ± 0.03 vs. 0.84 ± 0.06). When being accompanied by a confederate, the Low D group up-regulated their implicit moral attitudes to approach the High D group (Fig. 1C).

3.2. Dispositional justice sensitivity (JSI)

The presence of a confederate did not modulate the overall JSI scores of explicit justice sensitivities (\( t_{39} = 0.33, P = .74 \)). The modulation of explicit justice sensitivities
varied as a function of the individual baseline JSI scores, which was derived from the alone condition. $\triangle$JSI (confederate – alone) was negatively correlated with the baseline JSI in the alone condition ($r = -0.58, P < .001$) (Fig. 2A). We divided the groups of relatively high and low JSI based on the median split of baseline JSI scores (Median = 177; Mean ± SD = 175.88 ± 35.04). We normalized the individual $\triangle$JSI caused by the presence of the confederate and performed a non-parametric one-sample-t-test (Wilcoxon signed rank tests) separately for these groups. The results showed that the group with lower baseline JSI (Low JSI: $n = 20$) showed a significant increase in $\triangle$JSI ($Z = 1.98, P = .048$), whereas the group with higher baseline JSI (High JSI: $n = 20$) showed a significant decrease modulated by the presence of a confederate ($Z = 2.49, P = .017$)(Fig. 2B). Neither the gender effect nor the related interaction was significant (all $P > 0.1$).

Given that the confederate presence did not cause a universal effect across participants and its modulation varied as a function of individual baseline JSI, we consequently conducted a mixed-design ANOVA with a within-subject variable (session: alone vs. confederate) and a between-subject variable (group: High vs. Low JSI). There was an interaction between group and session ($F_{1,38} = 8.95, P = .005, \eta_p^2 = 0.191, (1-\beta)\sim100\%$). Post hoc analyses revealed that the presence of a confederate
increased JSI scores for the Low JSI group (alone vs. confederate: 149.45 ± 5.12 vs. 162.95 ± 5.83), whereas decreased JSI scores for the High JSI group (202.3 ± 2.12 vs. 191.75 ± 5.83) (Fig. 2C). Neither the gender effect nor the related interaction was significant ($P > 0.1$). There was no significant relationship between mIAT and JSI in any pairs (all $P > 0.1$).

3.3. Subjective ratings

For subjective ratings of morally-laden behaviors, a repeated ANOVA analysis comprising the within-subject factor as session (alone vs. confederate), was conducted for harming and helping actions, respectively. Blame ratings for harming actions ranged from 2.7 to 7 (mean ± SD: 4.3 ± 0.62). Praise ratings for helping actions ranged from 1.14 to 7 (3.86 ± 0.98). The presence of a confederate did not affect moral evaluation for either harming ($F_{1, 39} = 0.01, P = .91, \eta^2_p < 0.001$) or helping actions ($F_{1, 39} = 2.90, P = .10, \eta^2_p = 0.07$). Neither the gender effect nor the related interaction was significant ($P > 0.1$).

3.4. fMRI data

3.4.1. Voxel-wise analysis
The confederate as compared with the alone session (confederate – alone), irrespective of action, was not associated with any significantly increased activity, but lead to decreased responses in the supplementary motor area (SMA), dIPFC, dmPFC, aIN, vmPFC, amygdala, and pSTS/TPJ (Fig. 3). Irrespective of the sessions (alone or confederate), simulating harming relative to helping another was associated with significant responses in the aIN, amygdala, dmPFC, and pSTS/TPJ. The reverse contrast, helping relative to harming another, was associated with significant responses in the vmPFC and dIPFC (Table 2). Neither the gender effect nor the related interaction was significant (all \( P > 0.1 \)).

3.4.2. ROI results

To examine the association between the confederate effect found in the behavioral level of analysis, and the neural responses related to the main effect of confederate, the D and/or JSI scores were computed as a continuous variable related to the neural activity with FWE rate at \( P < .05 \). Step-down Holm–Bonferroni correction was used to control the familywise error to counteract the problem of multiple comparisons in the reported results. The modulated activities in the amygdala (\( r = 0.41, P = .008 \)) and aIN (\( r = 0.38, P = .019 \)) by the presence of a confederate were significantly correlated with the mIAT changes across the two sessions (\( \triangle D: \)).
confederate – alone) (Fig. 4A). Given that the presence of a confederate modulated the mIAT based on individual baseline D scores, we conducted a mixed-design ANOVA with session (alone vs. confederate) and action (harming vs. helping) as within-subject variables and group (High D vs. Low D) as a between-subject variable. The amygdala had a main effect of group ($F_{1, 38} = 5.45, P = .025, \eta_p^2 = 0.125$) and a marginal interaction between group and session ($F_{1, 38} = 3.24, P = .08, \eta_p^2 = 0.078$). The aIN had an interaction between group and session ($F_{1, 38} = 4.22, P = .047, \eta_p^2 = 0.1$). The follow-up analyses indicated that the group effect in the amygdala and aIN had opposite directions, depending on the factor of session (Fig. 4B). The presence of a confederate increased the activities in the amygdala (alone vs. confederate: $-0.08 \pm 0.07$ vs. $0.1 \pm 0.07$) and aIN ($-0.18 \pm 0.13$ vs. $-0.05 \pm 0.11$) in the Low D group, whereas it decreased their activities in the High D group (amygdala: $0.2 \pm 0.07$ vs. $0.1 \pm 0.07$; aIN: $0.13 \pm 0.13$ vs. $-0.16 \pm 0.11$). The High D relative to Low D group showed stronger activity in the amygdala-and aIN during the alone session, whereas no such group difference was detected during the confederate session. Neither the gender effect nor the related interaction was significant (all $P > 0.1$).

On the other hand, the JSI changes across the two sessions ($\Delta$JSI: confederate – alone) were significantly correlated with the modulated activities in the dmPFC ($r =$
0.40, \( P = .016 \) and dlPFC \((r = 0.53, \ P = .001)\) in the presence of a confederate (Fig. 5A). Given that confederate presence modulated the JSI based on individual baseline JSI scores, we conducted a mixed-design ANOVA with session (alone vs. confederate) and action (harming vs. helping) as within-subject variables, and group (High JSI vs. Low JSI) as a between-subject variable. There was a main effect of action \( (F_{1, \ 33} = 4.76, \ P = .036, \ \eta_p^2 = 0.126) \) and a three-way interaction (session x action x group) \( (F_{1, \ 33} = 9.96, \ P = .003, \ \eta_p^2 = 0.232) \) in the dmPFC. The follow-up analyses indicated that the session x group interaction was significant for the harming actions \( (F_{1, \ 33} = 4.36, \ P = .044, \ \eta_p^2 = 0.117) \), but not for the helping actions \( (F_{1, \ 35} = 0.17, \ P = .68, \ \eta_p^2 = 0.005) \).

While simulating harming actions, the High relative to Low JSI group showed stronger activity in the dmPFC during the alone session \((1.03 \pm 0.36 \ vs. \ -0.18 \pm 0.35)\), whereas no such group difference was detected during the confederate session \((-0.28 \pm 0.42 \ vs. \ 0.44 \pm 0.41)\). As for the dlPFC, the ANOVA revealed an interaction between group and session \( (F_{1, \ 35} = 10.02, \ P = .003, \ \eta_p^2 = 0.223) \). The follow-up analyses indicated that the group effect in the dlPFC had opposite directions, depending on the factor of session (Fig. 5B). The presence of a confederate increased the dlPFC activity in the Low JSI group (alone vs. confederate: \(-0.23 \pm 0.16 \ vs. \ 0.21 \pm 0.24\)), whereas it decreased the activity in the High JSI group \((0.15 \pm 0.18)\).
vs. \(-0.39 \pm 0.26\). Neither the gender effect nor the related interaction was significant (all \(P > 0.1\)).

4. Discussion

The goal of the present study was to extend the existing knowledge regarding the way by which social presence modulates the neural mechanisms underlying moral decision-making by examining the confederate effect on implicit moral attitudes (mIAT) and explicit justice sensitivities (JSI). Here, we measured mIAT and JSI, and scanned the neural activity during mental simulation of helping or harming behaviors, when participants were either alone or when they were accompanied by someone else who was doing the same moral tasks with them (confederate). The results showed that both the JSI and mIAT, as well as the neural correlates of moral behavior, are modulated and, further, synchronized by the presence of others. Considering that observation by others is a powerful elicitor of prosocial behavior (Andreoni and Petrie, 2004; Bateson et al., 2015; Jung et al., 2018; Lee et al., 2018), we extend the current literature on this topic by demonstrating how individual factors regarding moral attitudes (i.e., mIAT and JSI) play a role in the confederate effect. As a consequence of such effect, individual differences concerning moral attitudes and the neural correlates underlying moral behavior become synchronized. The present
findings fit well with the evolutionary significance of morality, since morality as a set of norms to be sanctioned originated from the group’s consensus and cultural evolution (Ayala, 2010).

The presence of a confederate reduced aIN and amygdala activities for harming, and vmPFC and dIPFC activities for helping. The aIN is involved in the representation of ongoing and predicted feeling states in relation to the opinions others might hold of oneself (Campbell-Meiklejohn et al., 2010). Recent research found that the aIN activity was increased in participants subject to non-social tasks that require low-effort and intuition, but this activity was decreased when they engage in social trials, as mentalizing processes where favored in order to adjust individual strategies to the strategies of the group (Beyer et al., 2018). With regards to the confederate effect in the vmPFC and dIPFC for helping, this result might be partially explained by the bystander effect, which refers to a decrease in helping behavior by any individual in a group of people who simultaneously witness a person in need of assistance (Darley and Latane, 1968; Hortensius and de Gelder, 2014). Previous research has attributed this reduction in helping behavior to self-serving bias, which refers to the self-protective attitudinal position assumed by individuals when taking credit in tasks leading to success, or blaming others when such actions lead to failure (Campbell et
The self-serving bias mitigates feelings of rejection, as it lets people attribute responsibility to factors other than themselves (Li et al., 2013).

In regards to the mIAT (as measured by D scores), participants with lower scores on the mIAT enhanced their moral attitudes when they were in the presence of others. This change was driven by increased signals in the amygdala and aIN when they mentally simulated harming another person. When a confederate was introduced, the Low D group up-regulated their mIAT, as well as the hemodynamic activity in the amygdala and aIN, in a way that it approached the High D group’s in the alone condition. Conversely, the introduction of a confederate decreased amygdala and aIN activity in the High D group. That is, during the confederate condition, an increase in the amygdala and aIN activity in the Low D group, but a decrease in the High D group, made their group difference to turn non-significant. It is well known that the amygdala plays a critical role in interpreting socially salient stimuli (e.g., facial expressions, although most apparent for fearful expressions), and in emotion regulation and control (Phelps and LeDoux, 2005), whereas the aIN processes the saliency of the ‘self’ itself in such a social context (Perini et al., 2018). So it seems reasonable to propose that the mIAT performance may indicate the ability of the participants to synchronize this facet of social cognition, in such a way that it flexibly
adjusts in response to the continued challenges in need of social group adhesion, and with moral attitudes just falling along to what is convenient at the time. Another likely explanation is in line with the argument in regards to the so called Salience Asymmetry of the IAT. This argument posits that the IAT doesn’t really measure associations but rather salience of concepts, and with those concepts drawing in more attentional resources (more salient) requiring more performance time, e.g., bad words or non-words being more salient than good and actual words, thus those trials with bad or non-words yielding higher RT than those with good or normal words (Greenwald et al., 2005; Popa-Roch and Delmas, 2011; Rothermund and Wentura, 2001, 2004). Consequently, it is possible that the mIAT might measure the salience of immoral behaviors/words vs. moral behaviors/words, or even the salience of moral identity (Burke and Stets, 2009; Nosek et al., 2007), and that because of this property it also somehow elicits activation of the amygdala and aIN. Nevertheless, this enquiry cannot be currently answered in this study, but it does open a new venue for future research.

When it comes to justice sensitivity (JSI), a similar synchronizing effect could be seen as that observed with the mIAT, but this time in regards to the dmPFC and dIPFC. The High JSI group had stronger dmPFC and dIPFC activation relative to the Low JSI
When a confederate was introduced, the Low JSI group up-regulated their JSI scores, as well as the hemodynamic activity in the dmPFC and dIPFC, whereas the High JSI group showed the opposite pattern. Previous research has observed that justice sensitivities can predict neural responses in the dmPFC and dIPFC, as evaluating good behaviors requires the engagement of these prefrontal brain areas (Yoder and Decety, 2014a). Furthermore, the dIPFC is specifically implicated in preserving representations of behaviors and their consequences as a mean to assign culpability (Buckholtz and Marois, 2012). Additionally, studies have also yielded findings that the dmPFC and dIPFC participate in a dynamic mechanism which identifies and guides adaptive behavior to the demands of the context (Taren et al., 2011). Our results show that this adaptation is possible not only at an individual level, but it can also drive justice sensitivity synchronization between the individuals in a group.

It is important to note some limitations of the present study. Firstly, our sample was mainly comprised of Han Chinese college students. Prosocial behavior in Chinese culture has been previously seen as being dependent on social relations, hence that what is morally wrong or right varies depending of social context. This stands in contrast with western cultures, where right or wrong can be said to be objectively
defined, thus it’s the main driver of social behavior and does not change due to social circumstances (Bedford and Hwang, 2003). Consequently, it is possible that the effect of the presence of a confederate will be greater in our sample than it would be in a sample comprised of non-Han Chinese participants. Notwithstanding, previous research using samples from western populations have demonstrated that the presence of others, being it literal or figurative, did have an impact on behavior (Bateson et al., 2015; Darley and Latane, 1968). Secondly, the sample size might not appear to be large, nevertheless, it is comparable to that used in other research (Greenwald et al., 1998; Karpinski and Hilton, 2001).

5. Conclusion

Taken together, this study provides aid in the task of elucidating how the presence of an unrelated stranger has the potential to synchronize individual discrepancies regarding moral attitudes, as well as the neural underpinnings implicated in this type of modulation. The synchronization of brain regions –along with the subsequent interpersonal alignment of moral attitudes– when a confederate is present, might be the consequence of evolutionary adaptation, as access to social information and behavioral imitation without investing much time in the meaning of such actions helps spread effective behaviors within populations, at the same time that cognitive
labor and effort are minimized (Derex and Boyd, 2018). In this aspect, our findings echo that of other research that posits that prosocial behavior might not be necessarily driven by such things as feelings of guilt or even justice and morality, but rather by something as simple as social conformity (Stallen and Sanfey, 2015). Since the phenomenon of social facilitation was first described (Triplett, 1898), a cursory examination of the literature revealed inconsistent findings regarding how the presence of others affected physical and cognitive performance. In answer to this issue, studies yielded findings demonstrating that the presence of others could bring about facilitated or impaired performance depending on the type of task being performed (Bond and Titus, 1983). As the continued existence of a group may depend on the ability to adjust moral attitudes in response to challenges; group morality, often codified to regulate behavior within a community, develops shared concepts and beliefs. Thus, the presence of others affects one's performance to the extent it can even synchronize it, for better or worse. Whether the effect is positive or negative depends on whether the task is familiar or novel, the nature of the audience, and one's physiological responses.
Code and data availability

The data that support the findings of this study and the code used for data analysis are available upon reasonable request to the corresponding author.

Author contributions

C.C. and Y.C. conceived and conceptualized the study; C.C. wrote the code for the paradigm with input from Y.C.; C.C. implemented the experiments; C.C. performed all data analysis; C.C., R.M.M., and Y.C. wrote the manuscript and performed the subsequent reviews and editing.

Acknowledgments

We thank Ya-Ann Yang for assisting with data collection and Professor Jean Decety for the morally-laden scenarios used as the stimuli and over which the mIAT was built upon. Without their generous contribution such endeavor wouldn’t have been possible. The study was supported by the Ministry of Science and Technology (MOST 108-2410-H-010-005-MY3; 108-2636-H-038-001-), National Yang-Ming University Hospital (RD2019-003; 2020-003), Taipei Medical University (DP2-108-21121-01-N-03-03), and the Brain Research Center, National Yang-Ming University from The Featured Areas Research Center Program within the framework
of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan (108BRC-B501).

Conflicts of interest

None of the authors have conflicts of interest to declare.

References

Akitsuki, Y., Decety, J., 2009. Social context and perceived agency affects empathy for pain: an event-related fMRI investigation. Neuroimage 47, 722-734.

Andreoni, J., Petrie, R., 2004. Public goods experiments without confidentiality: a glimpse into fund-raising. Journal of Public Economics 88, 1605-1623.

Ayala, F.J., 2010. Colloquium paper: the difference of being human: morality. Proc. Natl. Acad. Sci. USA. 107 Suppl 2, 9015-9022.

Bateson, M., Robinson, R., Abayomi-Cole, T., Greenlees, J., O'Connor, A., Nettle, D., 2015. Watching eyes on potential litter can reduce littering: evidence from two field experiments. PeerJ 3, e1443.

Bedford, O., Hwang, k.-k., 2003. Guilt and Shame in Chinese Culture: A Cross-Cultural Framework from the Perspective of Morality and Identity.

Beyer, F., Sidarus, N., Fleming, S., Haggard, P., 2018. Losing Control in Social Situations: How the Presence of Others Affects Neural Processes Related to Sense of Agency.
Bieri, D., Reeve, R.A., Champion, G.D., Addicoat, L., Ziegler, J.B., 1990. The Faces Pain Scale for the self-assessment of the severity of pain experienced by children: development, initial validation, and preliminary investigation for ratio scale properties. Pain 41, 139-150.

Bond, C.F., Titus, L.J., 1983. Social facilitation: A meta-analysis of 241 studies. Psychol. Bull. 94, 265-292.

Buckholtz, J.W., Asplund, C.L., Dux, P.E., Zald, D.H., Gore, J.C., Jones, O.D., Marois, R., 2008. The neural correlates of third-party punishment. Neuron 60, 930-940.

Buckholtz, J.W., Marois, R., 2012. The roots of modern justice: cognitive and neural foundations of social norms and their enforcement. Nat. Neurosci. 15, 655-661.

Burke, P.J., Stets, J.E., 2009. Future Research, Identity Theory. Oxford University Press, Inc., pp. 197-221.

Bzdok, D., Schilbach, L., Vogeley, K., Schneider, K., Laird, A.R., Langner, R., Eickhoff, S.B., 2012. Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. Brain Struct. Funct. 217, 783-796.

Campbell, W.K., Sedikides, C., Reeder, G.D., Elliot, A.J., 2000. Among friends? An examination of friendship and the self-serving bias. Br. J. Soc. Psychol. 39 ( Pt 2), 229-239.
Campbell-Meiklejohn, D.K., Bach, D.R., Roepstorff, A., Dolan, R.J., Frith, C.D., 2010. How the opinion of others affects our valuation of objects. Curr. Biol. 20, 1165-1170.

Chen, C., Martínez, R.M., Chen, Y., Cheng, Y., 2020. Pointing fingers at others: The neural correlates of actor-observer asymmetry in blame attribution. Neuropsychologia 136, 107281.

Chen, H.-C., Li, W.-C., Chou, T.-A., Cho, S.-L., 2002. Measuring the self-exposure of Chinese people using the implicit association test. Psychological Testing 49, 217-234.

Darley, J.M., Latane, B., 1968. Bystander intervention in emergencies: diffusion of responsibility. J. Pers. Soc. Psychol. 8, 377-383.

Dasgupta, N., 2013. Implicit attitudes and beliefs adapt to situations: a decade of research on the malleability of implicit prejudice, stereotypes, and the self-concept. Advances in Experimental Social Psychology 47, 233-279.

Decety, J., Michalska, K.J., Akitsuki, Y., 2008. Who caused the pain? An fMRI investigation of empathy and intentionality in children. Neuropsychologia 46, 2607-2614.

Decety, J., Michalska, K.J., Kinzler, K.D., 2012. The contribution of emotion and cognition to moral sensitivity: a neurodevelopmental study. Cereb. Cortex 22,
Decety, J., Porges, E.C., 2011. Imagining being the agent of actions that carry different moral consequences: an fMRI study. Neuropsychologia 49, 2994-3001.

Derex, M., Boyd, R., 2018. Social information can potentiate understanding despite inhibiting cognitive effort. Sci. Rep. 8, 9980.

Ellemers, N., Pagliaro, S., Barreto, M., 2013. Morality and behavioural regulation in groups: a social identity approach. European Review of Social Psychology 24, 160-193.

Ernest-Jones, M., Nettle, D., Bateson, M., 2011. Effects of eye images on everyday cooperative behavior: a field experiment. Evolution and Human Behavior 32, 172-178.

Faul, F., Erdfelder, E., Lang, A.G., Buchner, A., 2007. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav. Res. Methods 39, 175-191.

Fischer, P., Krueger, J.I., Greitemeyer, T., Vogrincic, C., Kastenmuller, A., Frey, D., Heene, M., Wicher, M., Kainbacher, M., 2011. The bystander-effect: a meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. Psychol. Bull. 137, 517-537.

Forbes, C.E., Grafman, J., 2010. The role of the human prefrontal cortex in social
cognition and moral judgment. Annu. Rev. Neurosci. 33, 299-324.

Greene, J.D., Nystrom, L.E., Engell, A.D., Darley, J.M., Cohen, J.D., 2004. The neural bases of cognitive conflict and control in moral judgment. Neuron 44, 389-400.

Greenwald, A.G., McGhee, D.E., Schwartz, J.L., 1998. Measuring individual differences in implicit cognition: the implicit association test. J. Pers. Soc. Psychol. 74, 1464-1480.

Greenwald, A.G., Nosek, B.A., Banaji, M.R., Klauer, K.C., 2005. Validity of the salience asymmetry interpretation of the implicit association test: comment on Rothermund and Wentura (2004). J. Exp. Psychol. Gen. 134, 420-425; author reply 426-430.

Gubbins, E., Byrne, R.M.J., 2014. Dual processes of emotion and reason in judgments about moral dilemmas. Thinking & Reasoning 20, 245-268.

Guerin, B., 2009. Social facilitation. Cambridge University Press, Cambridge, UK.

Hortensius, R., de Gelder, B., 2014. The neural basis of the bystander effect--the influence of group size on neural activity when witnessing an emergency. Neuroimage 93 Pt 1, 53-58.

Jung, D., Sul, S., Lee, M., Kim, H., 2018. Social Observation Increases Functional Segregation between MPFC Subregions Predicting Prosocial Consumer Decisions. Sci. Rep. 8, 3368.
Kahane, G., 2015. Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. Soc. Neurosci. 10, 551-560.

Karpinski, A., Hilton, J.L., 2001. Attitudes and the Implicit Association Test. J. Pers. Soc. Psychol. 81, 774-788.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., Damasio, A., 2007. Damage to the prefrontal cortex increases utilitarian moral judgements. Nature 446, 908-911.

Lamm, C., Decety, J., Singer, T., 2011. Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. Neuroimage 54, 2492-2502.

Lee, M., Sul, S., Kim, H., 2018. Social observation increases deontological judgments in moral dilemmas. Evolution and Human Behavior 39, 611-621.

Li, P., Shen, Y., Sui, X., Chen, C., Feng, T., Li, H., Holroyd, C., 2013. The neural basis of responsibility attribution in decision-making. PLoS One 8, e80389.

Nosek, B.A., Bar-Anan, Y., Sriram, N., Axt, J., Greenwald, A.G., 2014. Understanding and using the brief Implicit Association Test: recommended scoring procedures. PLoS One 9, e110938.

Nosek, B.A., Greenwald, A.G., Banaji, M.R., 2007. The Implicit Association Test at Age 7: A Methodological and Conceptual Review, Social psychology and the
unconscious: The automaticity of higher mental processes. Psychology Press, pp. 265-292.

Pascual, L., Rodrigues, P., Gallardo-Pujol, D., 2013. How does morality work in the brain? A functional and structural perspective of moral behavior. Front. Integr. Neurosci. 7, 65.

Perini, I., Gustafsson, P.A., Hamilton, J.P., Kampe, R., Zetterqvist, M., Heilig, M., 2018. The salience of self, not social pain, is encoded by dorsal anterior cingulate and insula. Sci. Rep. 8, 6165.

Phelps, E.A., LeDoux, J.E., 2005. Contributions of the amygdala to emotion processing: from animal models to human behavior. Neuron 48, 175-187.

Popa-Roch, M.-A., Delmas, F., 2011. Evidence for the Salience Asymmetries producing IAT effects: Replication and Extension with a French-North African IAT. Revue internationale de psychologie sociale 24, 37-61.

Rothermund, K., Wentura, D., 2001. Figure-ground asymmetries in the Implicit Association Test (IAT). Z. Exp. Psychol. 48, 94-106.

Rothermund, K., Wentura, D., 2004. Underlying processes in the implicit association test: dissociating salience from associations. J. Exp. Psychol. Gen. 133, 139-165.

Schaich Borg, J., Lieberman, D., Kiehl, K.A., 2008. Infection, incest, and iniquity: investigating the neural correlates of disgust and morality. J. Cogn. Neurosci. 20,
1529-1546.

Schmitt, M., Baumert, A., Gollwitzer, M., Maes, J., 2010. The Justice Sensitivity Inventory: Factorial validity, location in the personality facet space, demographic pattern, and normative data. Springer, Germany, pp. 211-238.

Stallen, M., Sanfey, A.G., 2015. The neuroscience of social conformity: implications for fundamental and applied research. Front. Neurosci. 9, 337.

Taren, A.A., Venkatraman, V., Huettel, S.A., 2011. A parallel functional topography between medial and lateral prefrontal cortex: evidence and implications for cognitive control. J. Neurosci. 31, 5026-5031.

Triplett, N., 1898. The dynamogenic factors in pacemaking and competition. American Journal of Psychology 9, 507-533.

Yoder, K.J., Decety, J., 2014a. The Good, the bad, and the just: justice sensitivity predicts neural response during moral evaluation of actions performed by others. J. Neurosci. 34, 4161-4166.

Yoder, K.J., Decety, J., 2014b. Spatiotemporal neural dynamics of moral judgment: a high-density ERP study. Neuropsychologia 60, 39-45.

Young, L., Dungan, J., 2012. Where in the brain is morality? Everywhere and maybe nowhere. Soc. Neurosci. 7, 1-10.
Table Legends

**Table 1:** Demographic and descriptive statistics in the alone and confederate conditions (N = 40).

|                   | Alone condition |                 | Confederate condition |                 |
|-------------------|-----------------|------------------|-----------------------|------------------|
|                   | Mean  | SD   | Mean   | SD    | Mean   | SD    |
| Age (yrs)         | 23.5  | 2.2  | 23.5   | 2.2   |
| mlAT              | 0.60  | 0.23 | 0.78   | 0.27  |
| JSI               | 175.86| 35.04| 177.35 | 29.58 |
| Blame ratings     | 4.29  | 0.62 | 4.30   | 0.63  |
| Praise ratings    | 4.00  | 0.84 | 3.70   | 1.10  |

**Table 2:** Social influences on brain regions showing significant BOLD responses for moral behaviors. Pooled group results for all participants (n = 40). All clusters are significant at FWE-corrected $P < 0.05$ [thresholded at $P < .001$, cut-off, $t = 3.145$ (uncorrected) with a spatial extent threshold $k > 10$], except those marked with an asterisk, which are taken from predefined ROIs and significant at uncorrected $P < .05$.

|                   | MNI Coordinates | t-value  |
|-------------------|------------------|----------|
|                   | x    | y    | z    |       |
| **Harming > Helping** |     |      |      |       |
| Amygdala          | L    | -28  | -6   | -16   | 2.69* |
| aIN               | L    | -36  | 16   | -6    | 4.22  |
| Insula            | L    | -38  | 6    | -4    | 4.01  |
| pSTS/TPJ          | R    | 58   | -42  | 18    | 1.92* |
| ACC               | R    | 4    | 22   | 26    | 4.33  |
| dmPFC             | L    | -6   | 48   | 32    | 3.7   |
| SMA               | L    | -14  | 12   | 64    | 4.4   |
| **Helping > Harming** |     |      |      |       |
| vmPFC             | R    | 8    | 50   | -14   | 2.60* |
| vmPFC             | L    | -14  | 40   | -10   | 3.93  |
| Brain Region                  | Side | MNI Coordinates | t-value |
|------------------------------|------|-----------------|---------|
| dIPFC                        | R    | 46 34 24        | 3.27*   |
| Alone > Confederate          |      |                 |         |
| Amygdala                     | L    | -28 2 -16       | 2.44*   |
| vmPFC                        | R    | 4 34 -14        | 3.26*   |
| aIN                          | R    | 34 20 -4        | 4.23    |
| aIN                          | L    | -30 28 -2       | 2.54*   |
| pSTS/TPJ                     | L    | -54 -58 18      | 3.70    |
| pSTS/TPJ                     | R    | 56 -56 20       | 3.45*   |
| dIPFC                        | L    | -44 26 22       | 2.96*   |
| Superior frontal cortex      | R    | 38 36 28        | 3.76    |
| Superior frontal cortex      | L    | -22 28 44       | 3.92    |
| SMA                          | R    | 2 24 46         | 4.35    |
| Middle frontal gyrus         | R    | 36 12 48        | 4.08    |
| Confederate > Alone          |      |                 |         |
| N.S.                         |      |                 |         |

Abbreviations: R, Right; L, left; PFC, prefrontal cortex; pSTS/TPJ, posterior superior temporal sulcus/temporal parietal junction; aIN, anterior insula; SMA, supplementary motor area; dmPFC, dorsomedial prefrontal cortex; dIPFC, dorsolateral prefrontal cortex; vmPFC, ventromedial prefrontal cortex; vSTR, ventral striatum; ACC, anterior cingulate cortex.
Figure Legends

Fig. 1. The confederate effect on implicit moral attitudes (mIAT).

A. Scatterplot showing the individual baseline D scores of the mIAT plotted against the $\Delta D$ (confederate – alone). Lower baseline D scores were associated with more $\Delta D$ induced by the presence of a confederate ($r = -0.61$, $P < .001$).

B. The group of participants who had lower baseline mIAT scores than the median (Low D: $n = 20$) showed a marginally significant increase ($Z = 1.89$, $P = .059$), whereas the group with higher baseline mIAT (High D: $n = 20$) was not significantly modulated by the presence of a confederate ($Z = 0.99$, $P = .32$). *, $P < .05$.

C. An ANOVA on the mIAT scores revealed an interaction between group (High vs. Low D) and session (alone vs. confederate) ($F_{1, 38} = 5.57$, $P = .024$). The mIAT differences between High and Low D scores (0.79 ± 0.03 vs. 0.42 ± 0.03) were eliminated due to the presence of a confederate (0.84 ± 0.06 vs. 0.72 ± 0.06). The Low D group up-regulated their mIAT. Error bars indicate ± SEM. ***, $P < .001$.

Fig. 2. The confederate effect on explicit justice sensitivities (JSI).

A. Scatterplot showing the individual baseline JSI scores plotted against the $\Delta JSI$ (confederate – alone). Lower baseline JSI scores were associated with more $\Delta JSI$ induced by the presence of a confederate ($r = -0.58$, $P < .001$).
B. The group of participants who had lower baseline explicit justice sensitivities than the median (Low JSI: $n = 20$) showed a significant increase ($Z = 1.98$, $P = .048$), whereas the group with higher baseline JSI scores (High JSI: $n = 20$) showed a significant decrease due to the presence of a confederate ($Z = 2.49$, $P = .017$). *, $P < .05$.

C. An ANOVA on the JSI revealed an interaction between group (High vs. Low) and session (alone vs. confederate) ($F_{1, 38} = 8.95$, $P = .005$). The presence of a confederate increased JSI scores for the Low JSI group (alone vs. confederate: $149.45 \pm 5.12$ vs. $162.95 \pm 5.83$), whereas decreased JSI scores for the High JSI group ($202.3 \pm 2.12$ vs. $191.75 \pm 5.83$). Error bars indicate $\pm$ SEM. **, $P < .01$; ***, $P < .001$.

**Fig. 3.** The confederate effect modulates BOLD activities elicited by mental simulation of moral behaviors. Distinct brain networks were involved in simulating harming and helping others. Harming was associated with increased BOLD signal in amygdala and insula, while helping recruited the ventromedial (vmPFC) and dorsolateral prefrontal cortex (dlPFC). Interestingly, the presence of a confederate reduced the neural responses in these regions. *, $P < .05$.

**Fig. 4.** The presence of a confederate reconciles group differences in implicit moral
attitudes and BOLD activities for moral behaviors.

A. The presence of a confederate modulated the amygdala ($r = 0.41, P = .008$) and aIN ($r = 0.38, P = .019$) activities and varied as a function of the changes in implicit moral attitudes across sessions ($\triangle D$: confederate – alone).

B. The group effect had opposite directions, depending on the factor of session. The presence of a confederate increased the activities in the amygdala (alone vs. confederate: $-0.08 \pm 0.07$ vs. $0.1 \pm 0.07$) and aIN ($-0.18 \pm 0.13$ vs. $-0.05 \pm 0.11$) in the Low D group, whereas it decreased their activities in the High D group (amygdala: $0.2 \pm 0.07$ vs. $0.1 \pm 0.07$; aIN: $0.13 \pm 0.13$ vs. $-0.16 \pm 0.11$). *, $P < .05$.

Fig. 5. The presence of a confederate reconciles group differences in justice sensitivities and BOLD activities for moral behaviors.

A. The presence of a confederate modulated the dmPFC ($r = 0.40, P = .016$) and dIPFC ($r = 0.53, P = .001$) activities and varied as a function of the JSI changes across the two sessions ($\triangle JSI$: confederate – alone).

B. The group effect had opposite directions, depending on the factor of session. The High relative to Low JSI group showed stronger activity in the dmPFC during the alone session ($1.03 \pm 0.36$ vs. $-0.18 \pm 0.35$), whereas no such group difference was detected during the confederate session ($-0.28 \pm 0.42$ vs. $0.44 \pm 0.41$). The presence
of a confederate increased the dIPFC activity in the Low JSI group (alone vs. confederate: \(-0.23 \pm 0.16\) vs. \(0.21 \pm 0.24\)), whereas it decreased the activity in the High JSI group \((0.15 \pm 0.18\) vs. \(-0.39 \pm 0.26\)). *, \(P < .05\).
