Current Biology

Article

Coercion Changes the Sense of Agency in the Human Brain

Highlights

- Responsibility for action is a key feature of human societies
- It depends on association between actions and outcomes in the brain
- Claims of reduced responsibility are sometimes based on "only obeying orders"
- Two experiments suggest coercion can reduce implicit measures of sense of agency

Authors

Emilie A. Caspar, Julia F. Christensen, Axel Cleeremans, Patrick Haggard

Correspondence

p.haggard@ucl.ac.uk

In Brief

Acting under coercion modifies the subjective experience of being the author of an action, reducing the perceived temporal association between actions and outcomes. Caspar et al. show that the neural processing of action outcomes under coercion more closely resembles situations of passive movement than actions carried out intentionally.

Caspar et al., 2016, Current Biology *26*, 1–8
March 7, 2016
http://dx.doi.org/10.1016/j.cub.2015.12.067

CellPress

Current Biology
Article

Coercion Changes the Sense of Agency in the Human Brain

Emilie A. Caspar,[1,2] Julia F. Christensen,[2] Axel Cleeremans,[1] and Patrick Haggard[2,*]
[1]Consciousness, Cognition, and Computation Group (CO3), Center for Research in Cognition and Neurosciences (CRCN), ULB Neuroscience Institute (UNI), Université libre de Bruxelles (ULB), Avenue F.D. Roosevelt 50, CP191, 1050 Brussels, Belgium
[2]Institute of Cognitive Neuroscience, University College London (UCL), Queen Square 17, London WC1N 3AR, UK
*Correspondence: p.haggard@ucl.ac.uk
http://dx.doi.org/10.1016/j.cub.2015.12.067

SUMMARY

People may deny responsibility for negative consequences of their actions by claiming that they were "only obeying orders." The "Nuremberg defense" offers one extreme example, though it is often dismissed as merely an attempt to avoid responsibility. Milgram's classic laboratory studies reported widespread obedience to an instruction to harm, suggesting that social coercion may alter mechanisms of voluntary agency, and hence abolish the normal experience of being in control of one's own actions. However, Milgram's and other studies relied on dissembling and on explicit measures of agency, which are known to be biased by social norms. Here, we combined coercive instructions to administer harm to a co-participant, with implicit measures of sense of agency, based on perceived compression of time intervals between voluntary actions and their outcomes, and with electrophysiological recordings. In two experiments, an experimenter ordered a volunteer to make a key-press action that caused either financial penalty or demonstrably painful electric shock to their co-participant, thereby increasing their own financial gain. Coercion increased the perceived interval between action and outcome, relative to a situation where participants freely chose to inflict the same harms. Interestingly, coercion also reduced the neural processing of the outcomes of one's own action. Thus, people who obey orders may subjectively experience their actions as closer to passive movements than fully voluntary actions. Our results highlight the complex relation between the brain mechanisms that generate the subjective experience of voluntary actions and social constructs, such as responsibility.

INTRODUCTION

In Milgram's classic experiments on obedience [1, 2], an experimenter ordered volunteer participants to inflict allegedly painful shocks to a third party. These studies focused on participants' readiness to conform to authority and obey coercive instructions to perform harmful actions. Interestingly, participants' subjective experience in such situations has not been systematically investigated, even though the legal defense of "only obeying orders" implies a loss of voluntary agency with coercion.

Sense of agency refers to the subjective experience of controlling one's actions, and, through them, external events. Explicit reports of perceived agency are modulated by numerous biases [3], notably social desirability and cognitive dissonance effects [4]. For example, individuals coerced into harmful actions might report reduced sense of agency for secondary gain, such as avoiding blame or punishment. Implicit measures may provide more direct access to the cognitive mechanisms underlying sense of agency, since these measures are less affected by task demands and social factors such as desirability. Here we used the perceived compression of time between a voluntary action and its outcome [5] as an appropriate implicit marker of sense of agency, and we investigated how coercion influenced this measure. Action-outcome intervals are perceived as shorter for intentional actions than for unintended actions such as passive movements [5, 6]. Therefore, if coercion indeed reduces the core experience of agency, interval estimates should be longer in the coercive than in the free-choice condition.

In a first experiment, participants were tested in pairs. They took turns being "agent" and "victim," ensuring reciprocity. In a first group of participants, the agent could freely choose on each trial to increase her own remuneration by taking money from the "victim" (financial harm). In a second, smaller group, the agent could freely choose to administer an electric shock to the "victim" (physical pain), again increasing her own remuneration. This free-choice condition was compared to a coercive condition, in which the experimenter stood next to the agent and ordered her before each trial whether to take money or not, or whether to shock the "victim" or not (see Figures 1 and 2).

RESULTS AND DISCUSSION

Experiment 1: Results

No participants withdrew from the experiment, and none reported any distress either after testing or at follow-up. In the financial harm group, agents freely chose to take money from the "victim" in 33.97/60 trials (95% confidence interval [CI] = 29.07–38.88, min 0, max 60). In the physical pain group, agents freely chose to give painful electric shocks to the "victim" in 31.37/60 trials (95% CI = 24.96–37.78, min 6, max 60). In

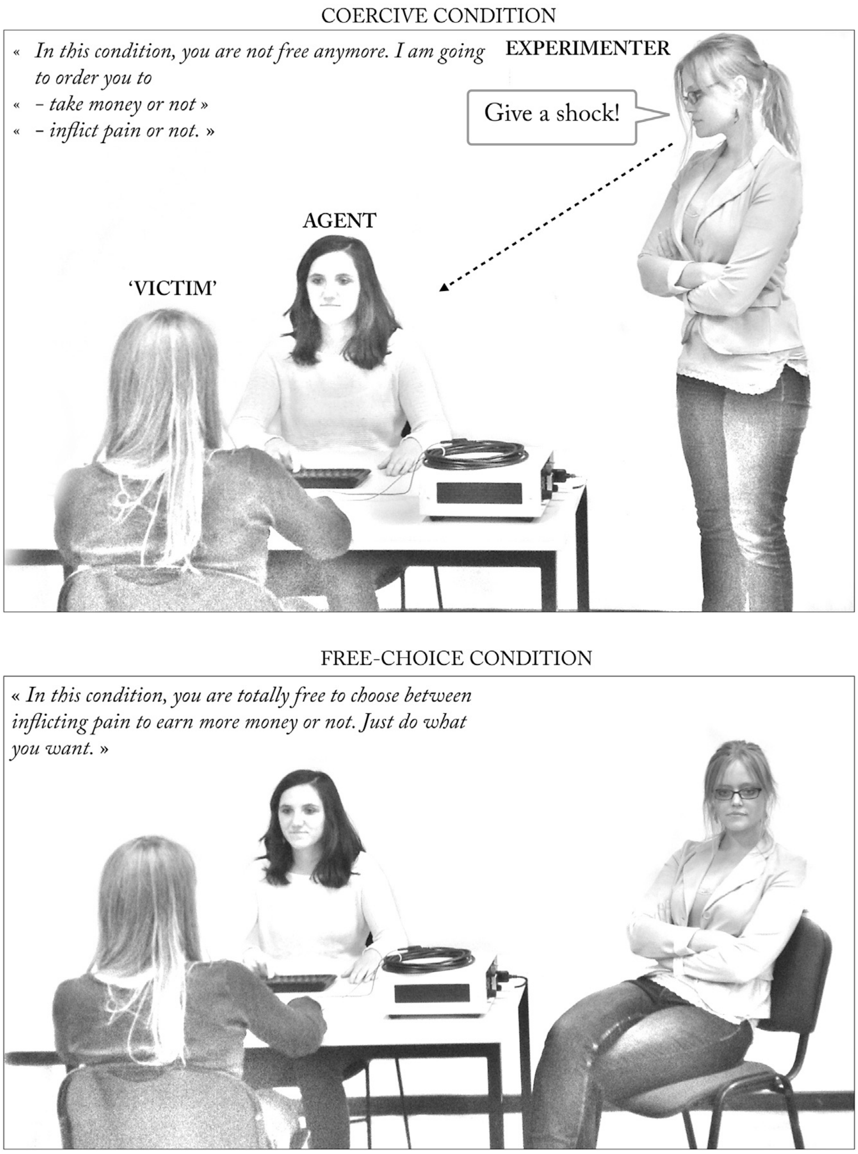

Figure 1. Experimental Setup
Schematic representation of the coercive condition (top) and the free-choice condition (bottom). In this condition, the experimenter looked elsewhere. In the coercive condition, the experimenter ordered the agent at each trial either to take money from her co-participant (financial harm group) or to deliver a shock (physical pain group). The experimenter stood next to the agent and looked at her throughout the whole condition.

addition, our free-choice condition captured key features of interpersonal choice, such as social reciprocity. In particular, experiencing pain as a "victim" guided subsequent free choices whether to inflict pain on one's co-participant. Regression analysis showed that, within the subgroup of participants who were first "victims" and then agents, participants who initially received high numbers of shocks as "victim" subsequently gave more shocks ($t(9) = 4.776$, $p = 0.001$, $R^2 = 0.860$). Such vindictive behavior is consistent with previous reports in economic games [7].

We analyzed agents' interval estimates using ANOVA, with condition (free choice, coercive) and outcome (harm, no harm) as within-subject factors, and group (financial harm, physical pain) as a between-subjects factor. The main effect of condition was significant ($F(1,50) = 22.740$, $p < 0.001$, $\eta^2_{partial} = 0.313$), with coercion leading to longer interval estimates than free choice (437 ms, 95% CI = 399–475, and 370 ms, 95% CI = 338–402, respectively; see Figure 4). There was a main effect of group, with lower interval estimates for the financial harm group than for the physical pain group ($F(1,50) = 6.042$, $p = 0.017$, $\eta^2_{partial} = 0.108$) but no evidence for any interaction between group and condition ($F(1,50) = 0.073$, NS). The main effect of outcome was not significant ($p > 0.3$). The interaction condition × outcome was not significant ($p = 0.099$; for full ANOVA table and further results, see the Supplemental Experimental Procedures). Interestingly, there was a significant interaction between outcome and group ($F(1,50) = 6.201$, $p < 0.02$, $\eta^2_{partial} = 0.110$). In the financial harm group, interval estimates decreased on trials when participants actually delivered harm (352 ms, 95% CI = 316–388) compared to when they did not (375 ms, 95% CI = 336–414; $t(34) = -2.699$, $p = 0.01$, Cohen's $d = 0.456$). However, in the physical pain group, this difference was not significant ($p > 0.2$). No other interactions with group were significant (all $ps > 0.3$). Importantly, there was no three-way interaction with choice condition—so we found no evidence that sense of agency varied specifically as a function of freely chosen outcomes. The effect of coercion was thus not related to whether or not agents delivered harm on any specific trial or to the content of any individual instruction, but was rather a contextual effect of receiving coercive instructions. This result also rules out explanations based on the attentional or arousing effects of harming others leading to altered time perception.

We then performed planned comparisons with our control conditions. Data for one participant in the active control condition was not available because of technical problems during testing. Planned comparisons showed that the free-choice condition produced shorter interval estimates than the active control condition ($t(56) = -2.809$, $p = 0.007$, Cohen's $d = 0.372$). Conversely, the coercive condition did not differ significantly from the passive condition ($p > 0.6$). These results suggest that voluntary actions made under coercion are experienced in some ways as if they were passive.

Please cite this article in press as: Caspar et al., Coercion Changes the Sense of Agency in the Human Brain, Current Biology (2016), http://dx.doi.org/10.1016/j.cub.2015.12.067

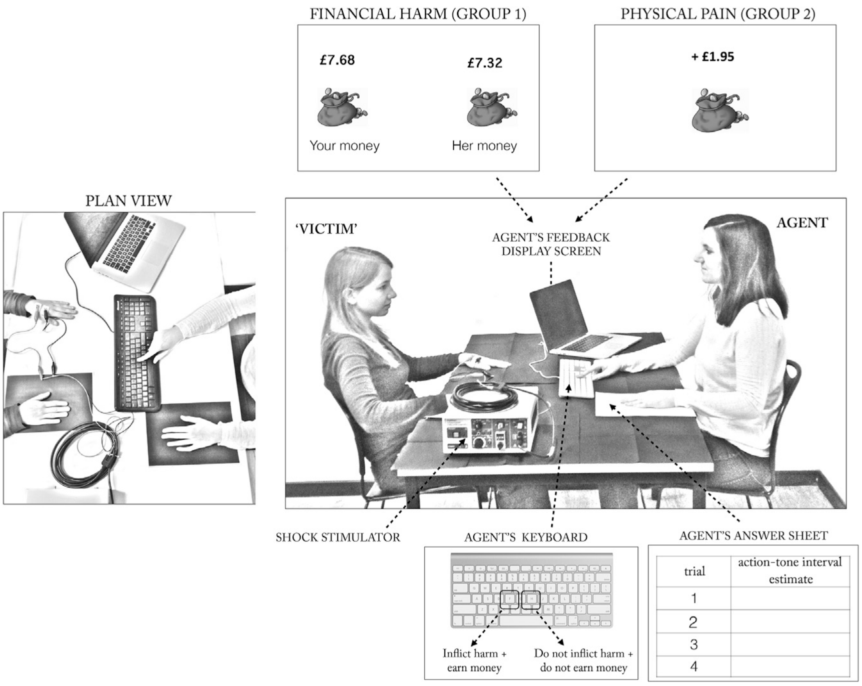

Figure 2. Schematic Representation of the Apparatus during the Experiment

The agent saw trial-by-trial feedback on the computer screen, whereas the "victim" did not. The agent pressed "F" on a keyboard to inflict harm and earn money or "H" not to inflict harm/earn money. Both the agent and the "victim" gave independent written estimates of action-tone intervals on an answer sheet. Electrodes connected to the stimulator were placed on the "victim's" left hand, which was clearly visible to the agent.

Because participants' free choices varied, we investigated whether any difference between free-choice and coercive conditions could simply reflect differences between these conditions in the number of harmful actions. We therefore added the difference between the number of harmful actions freely chosen by each participant and the number ordered by the coercive experimenter as a covariate. The covariate was not significant ($p > 0.7$), and the overall pattern of conclusions remained unchanged. Thus, our results reflect a specific effect of coercive *instruction* on the subjective experience of agency, occurring at the moment of voluntary action, rather than any difference between the contents of coercive instruction and free choices. In summary, when the agent was coerced, they experienced less agency than when they freely chose between the same options. This difference between coercion and choice did not interact with whether harm was actually inflicted or not.

Pre-session questionnaire responses allowed us to investigate whether sense of agency under coercion could be related to personality or trait empathy [8, 9]. We therefore explored whether personality and empathy measures were related to the "harm effect," i.e., the main effect difference between interval estimates associated with harmful actions and interval estimates associated with non-harmful actions. Questionnaire scores of trait empathy were positively and significantly related to the extent to which a harmful outcome event reduced participants' individual agency estimates (Table S1). More empathetic individuals showed a more dramatic reduction in sense of agency when their actions had harmful outcomes, compared to less harmful outcomes. Correlations with personality factors were generally weaker.

Additionally, we investigated whether the coercion effect, corresponding to the main effect difference between interval estimates associated with the coercive condition and interval estimates associated with the free-choice condition, correlated with personality and trait empathy (Table S2). Correlations were generally weaker than for the harm effect.

Experiment 1: Discussion

Why do people so readily comply with coercive instructions? This question remains central to historical [10] and psychological [11] investigations. The experience of agency under coercion has been surprisingly neglected in previous discussions, despite its obvious relation to personal responsibility. Here, we observed that being ordered to perform an action reduces the subjective experience of agency over the outcome in comparison with being free to choose between outcomes, as shown by reduced estimates of the temporal interval between action and outcome. Crucially, the effect of coercion was not related to whether harm actually occurred on any specific trial or to the content of any individual instruction (financial loss versus painful shock), but was rather a contextual effect of receiving coercive instructions. Sense of agency was previously shown to increase with the size of the "response space" of action choices [12]. A similar cognitive mechanism may explain why coercion both reduces the basic experience of agency and simultaneously increases compliance with instructions. "Only obeying orders" may not merely be a retrospective narrative of behavior, aimed at secondary gain such as blame avoidance, but may rather reflect a genuine difference in subjective experience of agency. Coercive instructions appear to induce a passive mode of processing in the brain compared to free choice between alternatives.

Experiment 2

In a second experiment, we investigated whether coercion changes brain activity, by focusing on electroencephalogram (EEG) potentials evoked by action outcomes. Filevich and colleagues [13] showed greater event-related potential (ERP) amplitudes for outcomes when participants freely chose an action compared with being instructed. We therefore predicted that coercive instructions should reduce outcome-evoked ERP amplitudes relative to free choices.

Experiment 2: Results

Behavioral Results

No participants withdrew from the experiment, and none reported any distress either after testing or at follow-up. Agents

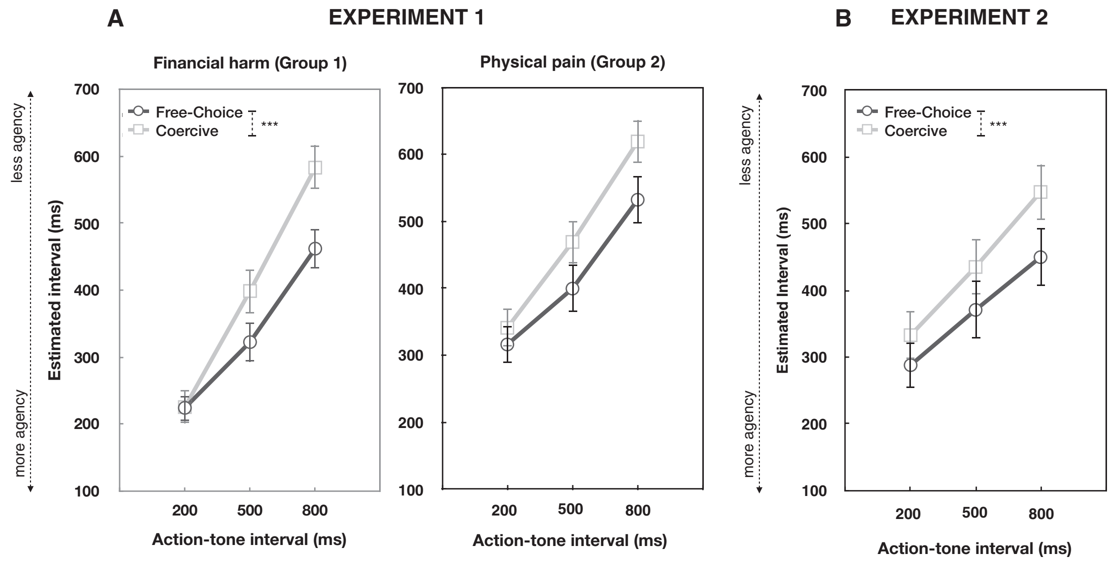

Figure 3. Interval Estimation Results

(A) Effects of coercion on interval estimates in experiment 1 in each group. The data for different action-tone intervals are shown to demonstrate interval estimation performance, but this factor was not central to our predictions. Error bars show SEs. Coercion consistently prolonged interval estimates. *** indicates a p value <0.001.

(B) Effects of coercion on interval estimates in experiment 2. *** indicates a p value <0.001.

freely chose to administer painful electric shocks to the "victim" in 18.75/60 trials (95% CI = 11.46–26.04, min 0, max 52). Regression analysis again showed modest support for vindictive behavior, as in experiment 1, with a trend for participants who served first as "victims" tending to choose to administer shocks in proportion to the number they had previously received ($t(9) = 1.681$, $p = 0.1$, $R^2 = 0.561$).

We also assessed how responsible participants felt during each condition by asking them to rate, in a post-session questionnaire, their responsibility as a percentage score in each condition. As expected, participants reported a higher degree of responsibility in the active condition (56.15%, SD = 39.70) than in the passive condition (17.92%, SD = 24.12; $t(19) = 3.792$, $p = 0.001$, Cohen's d = 0.847). Interestingly, they also reported feeling more responsible in the free-choice condition (86.85%, SD = 16.31) than in the coercive condition (34.80%, SD = 22.53; $t(19) = 9.832$, $p < 0.001$, Cohen's d = 2.19), but also than the active control condition (56.15%, SD = 39.70; $t(19) = 3.562$, $p = 0.002$, Cohen's d = 0.796). In addition, the degree of responsibility was higher in the coercive condition (34.80%, SD = 22.53) than in the passive control condition (17.92%, SD = 24.12; $t(19) = -2.699$, $p = 0.014$, Cohen's d = 0.603; see Figure 3).

Agents' interval estimates were analyzed using repeated-measures ANOVA, with condition (free choice, coercive) and outcome (harm, no harm) as within-subject factors. The main effect of condition was significant ($F(1,16) = 15.123$, $p = 0.001$, $\eta^2_{partial} = 0.486$), with free choice producing shorter interval judgments than coercion (366 ms, 95% CI = 288–444, and 424.5 ms, 95% CI = 351–498, respectively; see Figure 4). The main effect of outcome was not significant ($p > 0.8$), nor was the interaction condition × outcome ($p > 0.5$, for a full ANOVA table and further results, see the Supplemental Experimental Procedures). Planned comparisons with our control conditions showed that the passive and the coercive conditions did not differ ($p > 0.9$) and that the free-choice condition did not differ from the active condition ($p > 0.09$).

Event-Related Potentials

Standard ERP recording and processing methods were used (see the Supplemental Experimental Procedures).

We applied repeated-measures ANOVA to the auditory N1 amplitude [14], with the same ANOVA design used for interval estimates. The main effect of condition was significant ($F(1,16) = 8.009$, $p = 0.012$, $\eta^2_{partial} = 0.334$), with free choice producing more negative N1 amplitudes than coercion (-10.70 μv, 95% CI = -13.81 to -7.60, and -8.15 μv, 95% CI = -10.83 to -5.47, respectively). The main effect of outcome was not significant ($p > 0.1$), nor was the interaction condition × outcome ($p > 0.4$; for a full ANOVA table and further results, see the Supplemental Experimental Procedures; see Figure 5).

Interestingly, comparison between our active and passive control conditions showed a similar, but smaller, effect. Specifically, N1 amplitude was reduced in the passive condition (-9.99 μv, SD = 4.39) relative to the active condition (-11.11 μv, SD = 3.76; $t(19) = -1.814$, $p = 0.086$, Cohen's d = 0.405; see Figure 6). Thus, the psychological effect of coercion on sensory processing was similar to the physiological effect of subtracting voluntary motor commands from bodily movement [15]. Further, a direct comparison between the free-coercive difference, and the active-passive difference, expressed as the interaction term of a 2 × 2 ANOVA, showed that the modulation of sensory processing due to coercion was significantly stronger than the modulation by the voluntary motor command ($F(1,17) = 4.878$, $p = 0.041$, $\eta^2_{partial} = 0.223$). Direct planned comparisons between experimental and control conditions showed that the free-choice experimental condition did not differ from

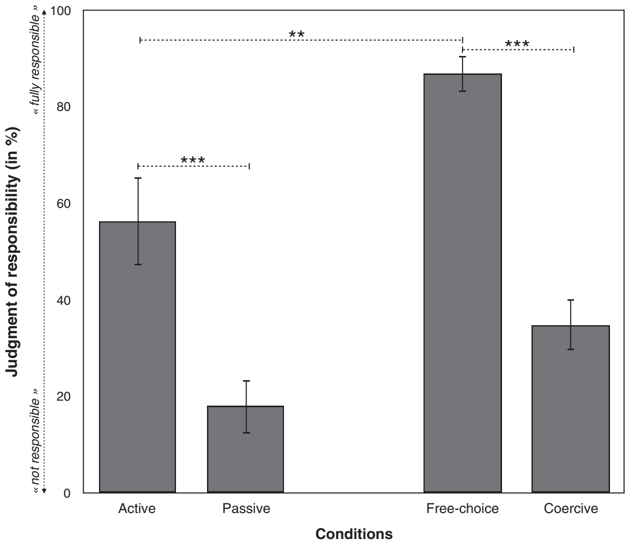

Figure 4. Judgments of Responsibility in Each Condition in Experiment 2

*** indicates a p value $\leq$0.001. ** indicates a p value between 0.001 and 0.01. Error bars show SEs.

the active control condition ($p > 0.6$). The auditory N1 was more pronounced in the passive condition (-9.99 μv, SD = 4.39) than in the coercive condition (-7.93 μv, SD = 4.85; $t(19) = 3.377$, $p = 0.003$, Cohen's d = 0.755).

Experiment 2: Discussion

The behavioral results replicated those of experiment 1. Receiving coercive orders again reduced the sense of agency over potentially harmful actions, according to our implicit measure based on intentional binding. Again, coercion produced an experience that more closely resembled passive movement than freely chosen voluntary action. This effect was again linked to the context of coercive instruction, rather than the actual harm resulting from any particular action. Explicit judgments of agency provided an important cross-validation of our implicit measures. Coercive orders thus influenced both explicit judgments of agency and the low-level subjective feeling of agency on which such judgments may be based [16]. The defense of ''only obeying orders'' is often treated with suspicion in law because of the clear secondary gain associated with denying responsibility. However, our result suggests that primary feelings and neurophysiological processing of agency are indeed reduced by coercion.

Analysis of the auditory N1 amplitude showed that coercion reduced processing of action outcomes compared to conditions in which participants freely chose what action to perform. This finding was again independent of whether the tone was accompanied by a painful electric shock to the co-participant or not. Importantly, the auditory tones were physically identical and equally predictable in all conditions. We suggest that coercive *contexts* produce an anticipatory reduction of sensory processing for action outcomes. This involves both downregulation of perceptual gains and temporal distancing. Indeed, the passive condition also displayed a reduction of the auditory N1 amplitude in comparison with the active condition, suggesting that the brain may treat consequences of one's actions under coercion as if they were passively triggered.

Interestingly, we found no evidence for sensorimotor attenuation of outcome processing [14, 17] when comparing either active versus passive movements or free versus coerced actions. However, the outcomes in our study occurred later than the short post-action window within which sensorimotor attenuation operates [18].

General Discussion

Issues of personal responsibility, moral action, and social influence are central to many accounts about human nature. Previous behavioral experiments have studied these issues using laboratory experiments [1]. However, the design and interpretation of those studies have been criticized. For example, Milgram's participants did not actually deliver pain, and pain responses of the third party were faked by an actor. It remains unclear whether Milgram's participants *really believed* they were delivering severe pain or whether they had some intuition that they were part of a simulation. In our design, participants acted reciprocally as agent and ''victim,'' both delivering and receiving harm. They knew from direct sensory experience how their actions would affect their co-participant. This experience demonstrably influenced their free choices. Moreover, the combination of money and pain in our experiments ensured that our participants were motivated by greed and fear, factors that may pervasively influence many human choices [19].

Legal, historical [10], and psychological [11] thought have all considered how obeying orders influences personal responsibility. Social constructs, including power [20] and authority [1], are often invoked. Using an implicit marker of sense of agency based on time perception, we showed that coercive instructions caused participants to *experience* less agency over the harmful outcomes of their actions. The results generalized over implicit and explicit measures of agency, and also over financial harm and physical pain, and they were also found on subsets of trials where no harm was actually delivered. Our results suggest that ''only obeying orders'' may not merely be a retrospective narrative of behavior, adopted for secondary gain such as mitigation, but may rather reflect a genuine difference in subjective experience of agency at the point of action itself. A previous study [12] reported that sense of agency decreased as the size of the ''response space'' [21] or alternative actions decreased. We suggest that this cognitive mechanism may also underlie the effects of coercion on sense of agency reported here.

Milgram reported that ''ordinary'' people frequently comply with coercive instructions [1]. Interestingly, our effects of coercion on sense of agency were quite general across individuals and were not strongly associated with particular personality traits or with empathy. This was not merely due to insensitivity of agency measures, since the effects of harm versus non-harm on sense of agency were higher in those with more empathic traits, as might be predicted. Rather, our result clearly suggests one reason why so many people can be coerced. Specifically, coercion may reduce the linkage that normally binds the experience of actions to their outcomes. Indeed, emotional distancing from distasteful outcomes of one's own necessary actions forms a specific part of training and professional culture in medicine [22] and in the military [23]. Training effects might

Please cite this article in press as: Caspar et al., Coercion Changes the Sense of Agency in the Human Brain, Current Biology (2016), http://dx.doi.org/10.1016/j.cub.2015.12.067

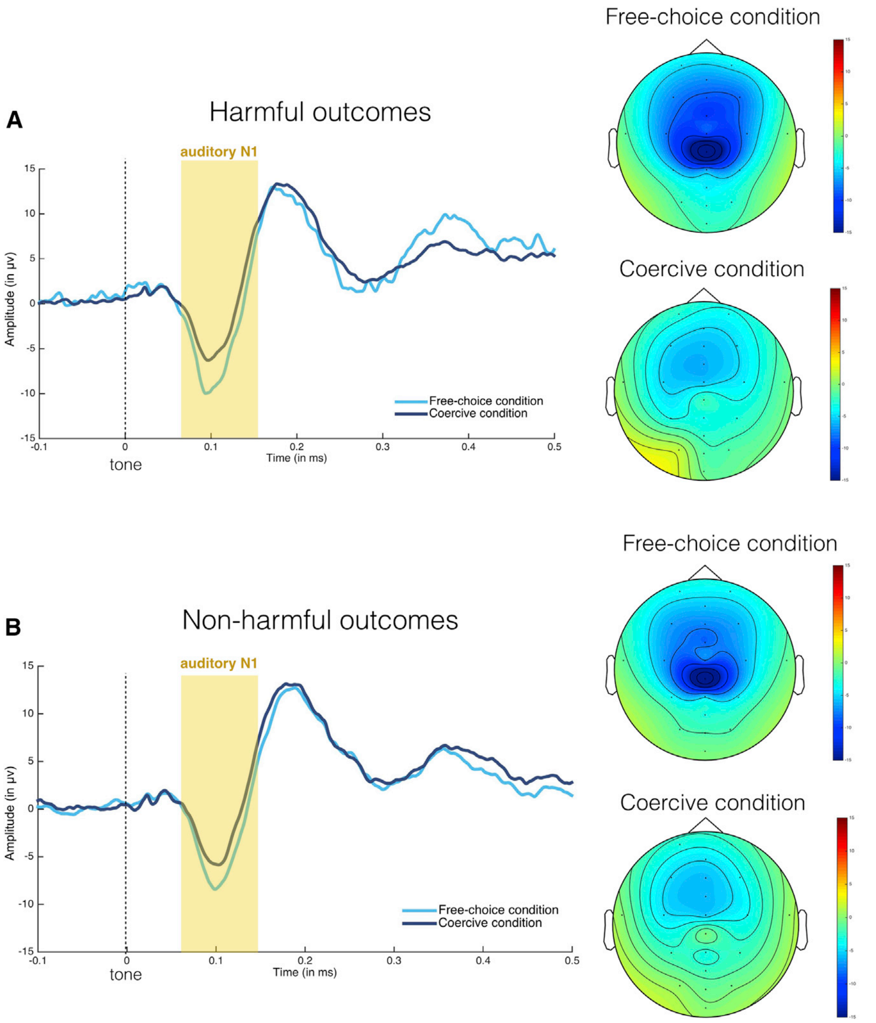

Figure 5. Neural Response to Outcome Tones: Experimental Conditions

Graphical representation of the auditory N1 amplitude in the free-choice (light blue) and the coercive (dark blue) conditions when (A) an electrical shock was delivered at the same time as the tone or (B) no electrical shock occurred. Topographical representations display the activity along the whole scalp.

also work in the opposite direction: learning the true valence of one's actions' outcomes might potentially make the sense of agency more resilient to the undermining effects of coercion.

We showed that acting under coercion deeply modifies the sense of being responsible for outcomes of one's actions. It also attenuates the neural processing of outcomes. Both results can be interpreted as a cognitive operation of "distancing," or reducing the linkage between one's own decision-making, action, and outcome. Our results may have profound implications for social and legal responsibility. Laws are culturally evolved rules for managing impact of individuals' behaviors on others. Laws must therefore engage with the psychological and neurocognitive mechanisms that drive individual actions. Our finding of reduced experience of agency under coercion does not legitimate Nuremberg-type defenses: society could still expect agents to try to resist evil [10, 24]. However, our results do suggest that people may indeed *experience* reduced agency at the point of being coerced to perform abhorrent actions. Clearly, society needs protection from harm, irrespective of whether the perpetrators experienced agency at the time of the act, or not. For example, the law argues that informed, rational agents should *know* they remain responsible for their actions, even if they have minimal *experience* of agency at the time of action. Further, the law could shift its focus away from those who obey orders toward those who *give* them, to prevent them from abusing a position that allows them to coerce others. Our research on the experience of agency highlights the fundamental link between law and cognitive neuroscience. The law has to engage with the human capacity to control action if it is to fulfil its function of allowing individuals to live together in societies.

EXPERIMENTAL PROCEDURES

Experiment 1

The principles of the 2013 Declaration of Helsinki were followed. The study was approved by University College London Research Ethics Committee (0847/006). All participants provided written informed consent prior to the experiment. No participant withdrew, and no participant reported distress at debriefing or at later follow-up.

Participants

Sixty right-handed student female participants were recruited in pairs and were paid £15–£25 for their participation. Only female participants were tested in order to control for potential effects of gender, both within the participant pairs, and also between participants and the (female) experimenters. Data from a number of standard questionnaires, including Big Five personality and trait empathy, were available from collection prior to participation. The protocol for matching participants in pairs stipulated that participants could not be relatives, friends, or from the same course or faculty. Thus, there was no particular relation between co-participants prior to the experiment. Data exclusion criteria were decided in advance of the experiment: failure to produce temporal intervals covarying monotonically with actual action-tone interval and any general failure to follow instructions. To identify participants for whom the action-tone intervals did not gradually increase with action-tone intervals, we performed linear trend analyses with contrast coefficients −1, 0, 1 for the three delays of the action-tone intervals (see the next section). Two participants were excluded due to non-significant linear trend. After this procedure, 39 participants remained in group 1 (mean age = 22.92, SD = 3.82) and 19 in group 2 (mean age = 22.79, SD = 3.63).

Materials and Procedure

On arrival at the experimental laboratory, participants read an information sheet about the experimental procedure and the aim of the experiment. The two co-participants signed their individual consent forms simultaneously, ensuring that they were both aware of the other's consent. Roles were assigned randomly so that one of the participants was told they were the agent and the other was the "victim." These roles were reversed for the second half of the experiment, making the procedure fully reciprocal. Participants sat at a table, face to face. An external, silent SODIAL Flexible Foldable USB keyboard

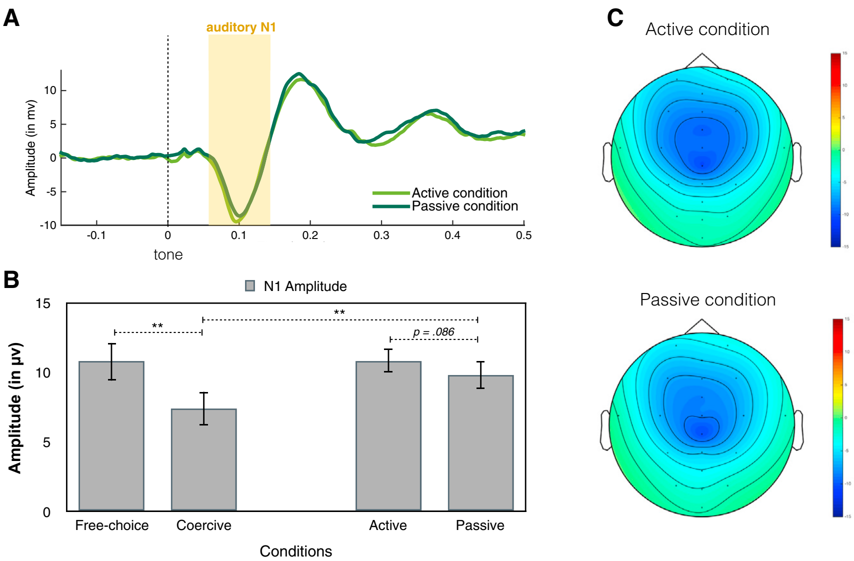

Figure 6. Neural Response to Outcome Tones: Control Conditions

(A) Auditory N1 amplitude in active (light green) and the passive (dark green) conditions.

(B) Mean amplitude of the auditory N1 in all conditions. ** indicates a significant difference (two-tailed, $p \leq 0.01$). Error bars show SEs.

(C) Topographical maps in active and passive conditions.

was placed between them, oriented toward the agent but visible to both. The experimental task ran on a computer, which was located on the agent's right side with the screen visible only to the agent (see Figure 2). The agent was instructed to press a key on the keyboard at a time she chose after the start of the trial, using the right index finger. This caused a tone to occur. The delay between key press and tone was set to vary randomly between 200, 500, and 800 ms. The participants' task was to estimate the delay between the key press and the tone. They were informed that the delay would vary randomly on a trial-by-trial basis, between 0 and 1,000 ms (they were reminded that 1,000 ms makes 1 s). Participants were also told to make use of all possible numbers between 1 and 1,000, as appropriate, avoiding restricting their answer space (e.g., not only use numbers between 1 and 100), and avoiding rounding (cf. [25]). Each participant received a paper sheet with 60 empty boxes for their time estimates in each condition of the task. Participants' answers were hidden from view of the other participant by a barrier, so as to avoid participants being biased by the other participant's answers.

There were two experimental and two control conditions. In the active control condition, the agent pressed the space key whenever she wanted. In the passive control condition, the experimenter pressed the agent's index finger down on the space bar, making sure to be unpredictable in her movements so as to minimize motor preparation in the agent. During these conditions, nothing was displayed on the computer screen. The predictions focused on two experimental conditions: free-choice and coercion. Participants were informed that they would both start with a specific amount of money (i.e., £20 for group 1 and £15 for group 2; this difference was due to the fact that participants in group 1 could lose money, but we needed to make sure participants would leave the experiment with the mandated minimum payment of £7.50/hr). In the free-choice condition, agents were instructed that they could freely choose to increase their remuneration for the experiment by delivering, or not delivering, a financial harm (group 1) or a physical pain (group 2) to the other participant, using the appropriate keys on the keyboard. They were told that they were totally free to choose how to act. The computer screen displayed which key press would be associated with which action (for instance, the "F" key for taking money/delivering a shock and the "H" key for refraining from taking money/delivering a shock). In group 1, the agents earned 5p each time they chose to inflict financial pain to the "victim," who then lost 5p. In group 2, the agents earned 5p each time they decided to deliver a painful electric shock to the "victim." They earned no money if they decided not to deliver a shock. During this condition, the experimenters did not look at the participants, but focused their attention on task irrelevant objects. In the coercive condition, the experimenters stood up next to the agent and ordered her, on each trial, to take money or not (group 1) or to administer a shock or not (group 2) to the "victim." The tone played after the key press was the same for both keys. In group 1, the "victim" did not know on each trial whether the agent had chosen to inflict financial harm or not and did not know which of the agent's response keys was mapped to financial harm—this information was available only on the agent's feedback display screen. Two new response keys ("J" and "L") were used when roles were reversed in group 1, so that former "victims" could not now simply repeat harm done to them by repeating the agent's previous key presses. Thus, group 1 was prevented from imitative behavior. In group 2, participants inevitably experienced the action selected by the agent on each trial, in the form of physical pain. Therefore, imitative behavior would become unavoidable: we thus decided to use the same key-press mappings throughout.

The gains and losses were displayed on the screen visible to the agent. In group 1, the agent saw two moneybags, 1 with her own money and 1 with the "victim's" money. Each time the agent inflicted financial harm, a coin was shown moving from the "victim's" bag to her bag, and the total amount of money increase was displayed. When the agent did not take money, no animations were displayed. In group 2, the agent only saw her own moneybag on the screen since the "victim" did not lose money, but instead received a painful shock to their left hand. The shock caused a twitch of the "victim's" hand that was readily visible to the agent.

Two experimenters participated in group 1, each testing half the sample. One experimenter instructed agents to take money 50/60 times. The other experimenter instructed agents to take money 30/60 times. This variation allowed some control over possible effects of experimenter's nastiness on participants' behavior and experiences under coercion. In group 2, both experimenters were simultaneously present, but one gave the coercive instructions. Both experimenters instructed agents to deliver shocks 30/60 times.

We used a partially randomized order of conditions. Participants performed the active or the passive control condition first, then the free-choice condition, then the coercive condition, and then the remaining control condition, either active or passive. We chose not to randomize free-choice and coercive conditions in order not to bias participants in the free-choice condition from previous experience in the coercive condition (e.g., attempting to match the coercive experimenter's instructions in their "free" choices). Participants went through the same four conditions twice, once as agent and once as "victim," that is, eight conditions in total. There were 60 trials per condition (20 trials at each action-tone delay, in randomized order), giving a total of 480 trials. Participants performed 240 actions as agents and observed 240 actions as "victims." The order of the conditions was the same within each pair.

Details of the painful stimulation and other measures are given in the Supplemental Experimental Procedures.

At the end of the experiment, participants were paid separately based on their earned financial gain during the experiment. For one dyad in group 1, the experimenter judged that the relation between the agent and the "victim" had become conflictual and hostile. The experimenter made an on-the-spot decision to pay both of these participants the same amount (£20), to reduce the possibility of subsequent distress or conflict.

Experiment 2

The principles of the 2013 Declaration of Helsinki were followed. The study was approved by the ethical committee of the Université libre de Bruxelles (018/2015). All participants provided written informed consent prior to the experiment. No participant withdrew, and no participant reported distress at debriefing or at later follow-up.

Participants

Twenty-two right-handed student female participants were recruited in pairs and paid €25–€31 for their participation. The same protocol for matching participants and the same data exclusion criteria than in experiment 1 were used. Two participants were excluded because no relation was found between perceived and actual action-shock intervals. After this procedure, 20 participants remained (mean age = 23.15, SD = 3.183).

Materials and Procedure

We used the same method as for the physical pain group in experiment 1, with modifications for EEG recording. Participants were instructed to wait a minimum of 2 s in a relaxed position before pressing a key, so as to obtain a consistent and noise-free baseline. Participants were further instructed not to move for up to 2 s after the tone. Participants first performed 30 trials in the active and the passive conditions, then 60 trials in the two experimental conditions, and then again 30 trials in the active and the passive conditions. These combinations of conditions were counterbalanced across participants. In order to have the same number of choices between the control and the experimental conditions, participants could choose between pressing "F" or "H" in the active condition. In the passive condition, the agent was asked to position two fingers (the index finger and the middle finger) on the two keys and the experimenter pressed down on one of the agent's fingers at an unpredictable moment in time. In the post-session questionnaire, we additionally asked participants to rate (from 0, "not responsible at all," to 100, "entirely responsible") how much they felt responsible in each condition. We also asked participants to rate how frequently they would have disobeyed if they could have (from -3, "almost never," to +3, "almost all the time"). In this experiment, the mean stimulation level selected by this procedure was 18.3 mA (SD = 6.7, pulse duration = 200 μs).

Source Data

The behavioral data reported in this paper have been published in Mendeley Data and are available at http://dx.doi.org/10.17632/322y43x9b7.1.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures and two tables and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2015.12.067.

AUTHOR CONTRIBUTIONS

E.A.C. developed the study concept. E.A.C. and P.H. created the study design, and J.F.C. and A.C. provided critical comments. E.A.C. and J.F.C. ran experiment 1. E.A.C. ran experiment 2. E.A.C. performed the data analysis and interpretation under the supervision of J.F.C., P.H., and A.C. E.A.C. drafted the manuscript, and J.F.C., P.H., and A.C. provided critical revisions. All authors approved the final version of the manuscript for submission.

ACKNOWLEDGMENTS

E.A.C. was supported by the FRS-F.N.R.S (Belgium). P.H. and J.F.C. were supported by AHRC grant L015145/1 to P.H. and ERC Advanced Grant HUMVOL. P.H. was additionally supported by an ESRC Professorial Research Fellowship. A.C. is a Research Director with the F.R.S.-FNRS (Belgium). This work was partly supported by BELSPO IAP grant P7/33 and by ERC Advanced Grant RADICAL to A.C.

Received: November 19, 2015
Revised: December 11, 2015
Accepted: December 23, 2015
Published: February 18, 2016

REFERENCES

1. Milgram, S. (1963). Behavioral study of obedience. J. Abnorm. Psychol. *67*, 371–378.
2. Milgram, S. (1974). Obedience to Authority: An Experimental View (Harper and Row).
3. Wegner, D.M. (2002). The Illusion of Conscious Will (The MIT press).
4. Bandura, A. (2006). Toward a psychology of human agency. Perspect. Psychol. Sci. *1*, 164–180.
5. Haggard, P., Clark, S., and Kalogeras, J. (2002). Voluntary action and conscious awareness. Nat. Neurosci. *5*, 382–385.
6. Engbert, K., Wohlschläger, A., and Haggard, P. (2008). Who is causing what? The sense of agency is relational and efferent-triggered. Cognition *107*, 693–704.
7. Singer, T., Seymour, B., O'Doherty, J.P., Stephan, K.E., Dolan, R.J., and Frith, C.D. (2006). Empathic neural responses are modulated by the perceived fairness of others. Nature *439*, 466–469.
8. Koban, L., Corradi-Dell'Acqua, C., and Vuilleumier, P. (2013). Integration of error agency and representation of others' pain in the anterior insula. J. Cogn. Neurosci. *25*, 258–272.
9. Lepron, E., Causse, M., and Farrer, C. (2015). Responsibility and the sense of agency enhance empathy for pain. Proc. Biol. Sci. *282*, 20142288.
10. Browning, C.R. (1998). Ordinary Men: Reserve Police Battalion 101 and the Final Solution in Poland (Harper Perennial).
11. Haslam, S.A., and Reicher, S. (2007). Beyond the banality of evil: three dynamics of an interactionist social psychology of tyranny. Pers. Soc. Psychol. Bull. *33*, 615–622.
12. Barlas, Z., and Obhi, S.S. (2013). Freedom, choice, and the sense of agency. Front. Hum. Neurosci. *7*, 514.
13. Filevich, E., Kühn, S., and Haggard, P. (2013). There is no free won't: antecedent brain activity predicts decisions to inhibit. PLoS ONE *8*, e53053.
14. Timm, J., SanMiguel, I., Keil, J., Schröger, E., and Schönwiesner, M. (2014). Motor intention determines sensory attenuation of brain responses to self-initiated sounds. J. Cogn. Neurosci. *26*, 1481–1489.
15. Wittgenstein, L. (1953). Philosophical Investigations (Blackwell).
16. Synofzik, M., Vosgerau, G., and Newen, A. (2008). Beyond the comparator model: a multifactorial two-step account of agency. Conscious. Cogn. *17*, 219–239.
17. Blakemore, S.J., Wolpert, D., and Frith, C. (2000). Why can't you tickle yourself? Neuroreport *11*, R11–R16.
18. Williams, S.R., Shenasa, J., and Chapman, C.E. (1998). Time course and magnitude of movement-related gating of tactile detection in humans. I. Importance of stimulus location. J. Neurophysiol. *79*, 947–963.
19. Coombs, C.H. (1973). A reparameterization of the prisoner's dilemma game. Behav. Sci. *18*, 424–428.
20. Foucault, M. (1977). Discipline and Punish: The Birth of the Prison (Allen Lane).
21. Nathaniel-James, D.A., and Frith, C.D. (2002). The role of the dorsolateral prefrontal cortex: evidence from the effects of contextual constraint in a sentence completion task. Neuroimage *16*, 1094–1102.
22. Tattersall, A.J., Bennett, P., and Pugh, S. (1999). Stress and coping in hospital doctors. Stress Health *15*, 109–113.
23. Johnsen, B.H., Laberg, J.C., and Eid, J. (1998). Coping strategies and mental health problems in a military unit. Mil. Med. *163*, 599–602.
24. Arendt, H. (1963). Eichmann in Jerusalem: A Report on the Banality of Evil (Penguin).
25. Caspar, E.A., Cleeremans, A., and Haggard, P. (2015). The relationship between human agency and embodiment. Conscious. Cogn. *33*, 226–236.

Current Biology, Volume *26*

Supplemental Information

Coercion Changes the Sense of Agency in the Human Brain

Emilie A. Caspar, Julia F. Christensen, Axel Cleeremans, and Patrick Haggard

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Table S1. Linear regression coefficients with empathy ratings as the independent variable and the "harm effect" as the dependant variable.

Questionnaires	R	F	Sig.
Interpersonal Reactivity Index - Total score	0.408	9.979	.003
IRI - Perspective taking	0.442	12.166	.001
IRI - Fantasy	0.493	16.079	.000
IRI - Empathic concern	0.481	15.037	.000
IRI - Personal distress	0.424	10.946	.002
Toronto Empathy Questionnaire	0.300	4.951	.031
Emotional Empathy	0.268	3.869	.055
Big Five Inventory	0.067	0.223	.639
BFI - Extraversion	0.351	7.717	.011
BFI - Agreeableness	0.162	1.352	.250
BFI - Conscientiousness	0.191	1.899	.174
BFI - Neuroticism	0.088	0.391	.534
BFI - Openness	0.336	6.366	.015

Interpersonal Reactivity Index, *IRI*, [S1]. *Perspective taking* = the tendency to spontaneously adopt the psychological point of view of others. *Fantasy* = taps respondents' tendencies to transpose themselves imaginatively into the feelings and actions of fictitious characters in books, movies and plays. *Empathic concern* = assesses 'other-oriented' feelings of sympathy and concern for unfortunate others. *Personal distress* = measures 'self-oriented' feelings of personal anxiety and unease in tense interpersonal settings [S1]. ***Toronto Empathy Questionnaire***, [S2]. ***Questionnaire measure of Emotional Empathy***, [S3]. ***Big Five Inventory***, BFI, [S4]. **Please see Experiment 1: Results.**

Table S2. Linear regression coefficients with empathy ratings as the independent variable and the "harm effect" as the dependant variable.

Questionnaires	R	F	Sig.
Interpersonal Reactivity Index - Total score	0.233	3.208	.079
IRI - Perspective taking	0.095	0.512	.477
IRI - Fantasy	0.200	2.324	.133
IRI - Empathic concern	0.219	2.809	.099
IRI - Personal distress	0.103	0.602	.441
Toronto Empathy Questionnaire	0.308	5.867	.019
Emotional Empathy	0.073	0.303	.584
Big Five Inventory	0.226	3.005	.639
BFI - Extraversion	0.129	0.953	.333
BFI - Agreeableness	0.126	0.904	.346
BFI - Conscientiousness	0.081	0.370	.545
BFI - Neuroticism	0.082	0.378	.541
BFI - Openness	0.220	2.853	.097

Please see Experiment 1: Results.

EXPERIMENT 1

All participants completed a short post-session questionnaire assessing how they had felt during the experiment. The first item assessed how bad they had felt to take money from/give shocks to the other participant and the second how sorry they were (answers were provided on 7-point Likert scales; "-3"; not very bad; "3"; very bad). In an open question, participants were invited to describe in a couple of words what they had felt during the experiment or what their thoughts were about this experiment. These descriptions confirmed that the participants were aware that their actions causal financial harm or physical pain to their co-participant, and that they generally "felt bad" about doing so. In Group 2, participants were additionally asked how much they had looked at the other person's face during the experiment ("-3"; not much; "3"; a lot) and how much pain they thought the other person had felt ("-3"; none at all; "3"; a lot). To assess participants' general time perception, they were asked to estimate the length of each of the 8 conditions in minutes.

In Group 2, pain was delivered using a constant current stimulator (Digitimer DS7A) connected to two electrodes placed on the back of victims' left hand, visible to the agent. Participants' individual pain threshold was determined for both participants after they had signed the consent form, before starting the experiment. This threshold was determined by increasing stimulation in steps of 1 mA. We approximated an appropriate threshold by asking a series of questions about their pain perception during the calibration (1. « *Is it uncomfortable?* » - 2. « *Is it painful?* » - 3. « *Could you cope with a maximum of 100 of these shocks?* » - 4. « *Could we increase the threshold?* » - 5. « *On a scale from 0 to 10, where 0 is not painful at all and 10 is the worst possible pain you can imagine; how would you rate this stimulus?* »). With this procedure, we ensured that participants were aware of the pain they were going to inflict to the other participant, and were willing to experience themselves. Participants kept the electrodes on their hand during the whole experiment, but only the victim's electrodes were connected to the electrical stimulator. The left hand of the victim was placed on the table and visible to the agent. When roles were reversed, we briefly re-calibrated the pain threshold of the new victim by increasing the stimulation again from 0 in steps of 3 mA up to the previously determined threshold, to confirm that the initial estimate was still appropriate, and to allow re-familiarisation. When this re-calibration disagreed with the initial re-calibration (7/20 participants), the second calibration was used as the final value. The mean stimulation level selected by this procedure was 21.8 mA (SD=7.74, pulse duration: 200 μs).

- **EFFECT TABLE OF REPEATED-MEASURES ANOVA**

Factors: Condition (Free, Coercive) x Outcome (Harm, No harm) as within-subjects factors and Group (Financial harm, Physical pain) as between-subjects factor.
Dependent variable: mean interval judgment.

Within-subjects effects

	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Condition	203557.810	1	203557.810	22.740	.000	.313
Condition*Group	652.738	1	652.738	.073	.788	.001
Error (Condition)	*447578.019*	*50*	*8951.560*			
Outcome	1577.470	1	1577.470	.739	.394	.015
Outcome*Group	13232.140	1	13232.140	6.201	.016	.110
Error (Outcome)	*106699.842*	*50*	*2133.997*			
Condition*Outcome	4790.083	1	4790.083	2.819	.099	.053
Condition*Outcome*Group	68.147	1	68.147	.040	.842	.001
*Error (Condition*Outcome)*	*84960.740*	*50*	*1699.215*			

Between-Subjects Effects

	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	29778017.5	1	29778017.5	633.836	.000	.927
Group	283857.495	1	283857.495	6.042	.017	.108
Error	*2349033.50*	*50*	*49-6980.670*			

Tables of marginal means

	Mean	Standard Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Grand Mean	403.303	16.019	371.127	435.478
Condition				
Free-choice	369.958	15.929	337.964	401.952
Coercive	436.647	18.902	398.681	474.614
Outcome				
Harm	400.367	15.753	368.727	432.007
No Harm	406.238	16.982	372.128	440.348
Group				
Financial Harm	363.927	18.319	327.132	400.721
Physical Pain	442.679	26.285	389.884	495.473
Condition*Outcome				
Free-choice - Harm	372.138	16.665	338.665	405.611
Free-choice – No Harm	367.778	16.283	335.073	400.483
Coercive - Harm	428.597	18.721	390.995	466.199
Coercive – No Harm	444.698	20.309	403.905	485.490
Group*Condition				
Financial Harm - Free-choice	332.470	18.215	295.884	369.056
Financial Harm - Coercive	395.383	21.616	351.966	438.800
Physical Pain - Free-choice	407.446	26.136	354.950	459.942
Physical Pain - Coercive	477.912	31.016	415.615	540.208
Group*Outcome				
Financial Harm - Harm	352.490	18.014	316.308	388.671
Financial Harm - No Harm	375.363	19.420	336.357	414.370
Physical Pain - Harm	448.245	25.847	396.329	500.161
Physical Pain – No Harm	437.113	27.865	381.144	493.082
Group*Condition*Outcome				
Financial Harm				
Free-choice - Harm	326.758	19.057	288.480	365.036
Free-choice – No Harm	338.182	18.620	300.782	375.581
Coercive - Harm	378.221	21.408	335.221	421.220
Coercive – No Harm	412.545	23.225	365.897	459.193
Physical Pain				
Free-choice - Harm	417.517	27.345	362.594	472.441
Free-choice – No Harm	397.375	26.717	343.712	451.038
Coercive - Harm	478.973	30.718	417.274	540.671
Coercive – No Harm	476.851	33.324	409.917	543.784

- **EFFECT TABLE OF REPEATED-MEASURES ANCOVA**

Within-Subjects Effects

	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Condition	183321.079	1	183321.079	20.233	.000	.292
Condition*Harmful Actions	3614.992	1	3614.992	.399	.531	.008
Condition*Group	1965.384	1	1965.384	.217	.643	.004
Error (Condition)	*443963.027*	*49*	*9060.470*			
Outcome	666.195	1	666.195	.317	.576	.006
Outcome*Harmful Actions	3759.405	1	3759.405	1.789	.187	.035
Outcome*Group	7674.917	1	7674.917	3.653	.062	.069
Error (Outcome)	*102940.437*	*49*	*2100.825*			
Condition*Outcome	4528.629	1	4528.629	2.612	.112	.051
Condition*Outcome*Harmful actions	2.586	1	2.586	.001	.969	.000
Condition*Outcome*Group	52.256	1	52.256	.030	.863	.001
*Error (Condition*Outcome)*	*84958.154*	*49*	*1733.840*			

Between-Subjects Effects

	Type III Sum of Squares	Df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	28722955.5	1	28722955.5	603.185	.000	.925
Covariate (Harmful actions)	15712.207	1	15712.207	.330	.568	.007
Group	210932.409	1	210932.409	4.430	.040	.083
Error	*2333321.29*	*49*	*47618.802*			

EXPERIMENT 2

- **EFFECT TABLE OF REPEATED-MEASURES ANOVA**

Factors: Condition (Free, Coercive) x Outcome (Harm, No harm) as within-subjects factors and Group (Financial harm, Physical pain) as between-subjects factor.
Dependent variable: mean interval judgment.

Within-Subjects Effects

	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Condition	57684.393	1	57684.393	15.123	.001	.486
Error (Condition)	*61031.112*	*16*	*3814.444*			
Outcome	30.659	1	30.659	.017	.897	.001
Error (Outcome)	*21411.494*	*16*	*1775.718*			
Condition*Outcome	559.191	1	559.191	.323	.578	.020
*Error (Condition*Outcome)*	*27687.065*	*16*	*1730.442*			

Table of Marginal Means

	Mean	Standard Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Grand Mean	395.519	34.974	321.378	469.661
Condition				
Free-choice	366.394	36.801	288.380	444.407
Coercive	424.645	34.703	351.079	498.211
Outcome				
Harm	396.191	35.358	321.235	471.147
No Harm	394.848	35.332	319.947	469.749
Condition*Outcome				
Free-choice - Harm	372.138	16.665	338.665	405.611
Free-choice – No Harm	367.778	16.283	335.073	400.483
Coercive - Harm	428.597	18.721	390.995	466.199
Coercive – No Harm	444.698	20.309	403.905	485.490

- **EFFECT TABLE OF REPEATED-MEASURES ANOVA**

<u>Factors:</u> Condition (Free, Coercive) x Outcome (Harm, No harm) as within-subjects factors and Group (Financial harm, Physical pain) as between-subjects factor.
<u>Dependent variable:</u> auditory N1 amplitude.

Within-Subjects Effects

	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Condition	110.782	1	110.782	8.009	.012	.334
Error (Condition)	*221.306*	*16*	*13.832*			
Outcome	11.607	1	11.607	1.803	.198	.101
Error (Outcome)	*102.982*	*16*	*6.436*			
Condition*Outcome	6.116	1	6.116	.722	.408	.043
*Error (Condition*Outcome)*	*135.550*	*16*	*8.472*			

Table of Marginal Means

	Mean	Standard Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Grand Mean	-9.431	1.290	-12.166	-6.695
Condition				
Free-choice	-10.707	1.462	-13.807	-7.607
Coercive	-8.154	1.264	-10.834	-5.474
Outcome				
Harm	-9.018	1.384	-11.951	-6.084
No Harm	-9.844	1.267	-12.529	-7.158
Condition*Outcome				
Free-choice - Harm	-9.994	1.598	-13.382	-6.606
Free-choice – No Harm	-11.420	1.493	-14.586	-8.254
Coercive - Harm	-8.041	1.286	-10.768	-5.314
Coercive – No Harm	-8.268	1.383	-11.200	-5.335

SUPPLEMENTAL REFERENCES

[S1] Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology*, *10*, 85.

[S2] Spreng, R. N., McKinnon, M. C., Mar, R. A., & Levine, B. (2009). The Toronto Empathy Questionnaire: scale development and initial validation of a factor-analytic solution to multiple empathy measures. *Journal of Personality Assessment*, *91*(1), 62-71.

[S3] Mehrabian, A., & Epstein, N. (1972). A measure of emotional empathy. *Journal of Personality*, *40*(4), 525-543.

[S4] John, O. P., & Srivastava, S. (1999). *The Big-Five trait taxonomy: History, measurement, and theoretical perspectives*. In L. A. Pervin & O. P. John (Eds.), Handbook of personality: Theory and research (Vol. 2, pp. 102–138). New York: Guilford Press.

[S5] Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2010). FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational intelligence and neuroscience*, *2011*.