How social contexts affect cognition: mentalizing interferes with sense of agency during voluntary action

Nura Sidarus[1,2], Eoin Travers[3], Patrick Haggard[3], Frederike Beyer[4*]

[1] Institut Jean Nicod, Département d'Études Cognitives, Ecole Normale Supérieure, EHESS, CNRS, PSL University

[2]Department of Psychology, Royal Holloway University of London, Surrey, United Kingdom of Great Britain and Northern Ireland

[3] Institute of Cognitive Neuroscience, University College London

[4] Department of Biological and Experimental Psychology, Queen Mary University of London

* Corresponding author

Dr. Frederike Beyer

Department of Biological and Experimental Psychology

School of Biological and Chemical Sciences

Queen Mary University of London

Mile End Road

E1 4NS, London

Phone: +44 (0)20 7882 7183

Email: f.beyer@qmul.ac.uk

Abstract

Living in complex social structures, humans have evolved a unique aptitude for mentalizing: trying to understand and predict the behaviour of others. To date, little is known about how mentalizing interacts with other cognitive processes. "Sense of agency" refers to the feeling of control over the outcomes of one's actions, providing a precursor of responsibility. Here, we test a model of how social context influences this key feature of human action, even when action outcomes are not specifically social. We propose that in social contexts, sense of agency is affected by the requirement to mentalize, increasing the complexity of individual decision-making. We test this hypothesis by comparing two situations, in which participants could either consider potential actions of another person (another participant acting to influence the task), or potential failures of a causal mechanism (a mechanical device breaking down and thereby influencing the task). For relatively good outcomes, we find an agency-reducing effect of external influence only in the social condition, suggesting that the presence of another intentional agent has a unique influence on the cognitive processes underlying one's own voluntary action. In a second experiment, we show that the presence of another potential agent reduces sense of agency both in a context of varying financial gains or of losses. This clearly dissociates social modulation of sense of agency from classical self-serving bias. Previous work primarily focused on social facilitation of human cognition. However, when people must incorporate potential actions of others into their decision-making, we show that the resulting socio-cognitive processes reduce the individuals' feelings of control.

sense of agency; social context; mentalizing; outcome processing

Introduction

Humans live in highly complex cooperative social structures, a fact that is linked to the development of sophisticated mentalizing skills during recent evolution (Hare, 2011). Mentalizing can be defined as the cognitive processes associated with trying to understand and predict the behaviour of another agent in a social interaction. The evolution of the human brain appears directly driven by the need for such complex social cognition, with a wide-ranging network of neural structures (medial prefrontal cortex; temporo-parietal junction; temporal poles; precuneus) supporting mentalizing processes (Schurz et al., 2014). This would suggest that the mentalizing processes underlying social interaction have shaped other, non-social cognitive processes (Mercier & Sperber, 2011). In that case, consistent and characteristic interactions between mentalizing and non-social cognition should exist. However, the tasks used in much previous research on this topic often *assumed* this interaction, rather than directly test it – often requiring social cognition as an explicit element of the task. For example, when participants need to learn to predict another agent's behaviour, mentalizing is indeed related to better performance (Devaine et al., 2014).

Despite its generally adaptive value, we suggest that, in some contexts, mentalizing may have a deleterious effect on cognition and behaviour. A troubling example of how social context can impact individuals' behaviour is the "bystander effect" (Darley and Latane, 1968), in which the presence of other people reduces the likelihood that any one individual will act in an emergency situation, like someone needing help. This effect has been linked to the phenomenon of diffusion of responsibility (Bandura, 1991), whereby people feel less responsible for their own actions in

social contexts. We recently proposed that these effects are due to mentalizing processes interfering with decision-making and sense of agency (Beyer et al., 2017).

Sense of agency refers to the feeling of being in control of our actions and their outcomes, and is essential for attribution of responsibility (Frith & Haggard, 2018). Sense of agency is an essential feature of normal human behaviour, and has wide structuring effects on cognitive processes, from perception (Tsakiris & Haggard, 2005) to outcome evaluation (Bednark & Franz, 2014). It is understood as arising from monitoring one's own volitional control over a physical event. Models of motor control (Blakemore et al., 2002) have highlighted a role for detecting mismatches in the comparison between internal predictions of sensory feedback, given efferent motor commands, with observed sensory feedback. Recent frameworks have emphasised an integration of such sensory-motor signals with other relevant cues, such as contextual information, or information about the decision-making process (Chambon et al., 2014; Synofzik et al., 2013). Traditionally, sense of agency is measured as a non-social aspect of cognition, which depends on action-outcome contingencies in interactions of the individual with their environment (Wen, 2019). Yet, navigating the social world raises particular opportunities and challenges for individual agency.

Social contexts offer the opportunity of expanding one's agency by acting together with, or through, other agents. This can be supported by socio-cognitive processes, such as reflective mentalizing, or automatic mimicry. Interestingly, another view, akin to models of motor control, conceptualises social interaction as a feedback loop, between one's own actions and outcomes and that of other agents, which would serve to facilitate coordination, as well allow assessing one's control over the interaction partner (Wolpert et al., 2003). Yet, while this model addresses how one may come to feel a sense of control over the interaction partner's actions, it does not

address the question of how the interaction partner affects one's own sense of agency over non-social, environmental consequences of one's own behaviour. In fact, social interactions can also present challenges to monitoring one's own agency. Namely, they can introduce ambiguity as to which of two or more potential agents caused a given event. Several studies have tested the effect of social interaction on sense of agency, particularly in joint action (Bolt et al., 2016), or in situations in which control over events is objectively shared between participants (Li et al., 2011). Using experimental designs that prevent such ambiguity as to who caused a given outcome, our work has demonstrated a different challenge to sense of agency, as social contexts can also increase the complexity of individual decision-making (Beyer et al., 2017, 2018).

Previously, we have shown that the mere presence of another potential agent alters decision-making, and reduces sense of agency and outcome monitoring (Beyer et al., 2017). Interestingly, this agency-reducing effect of social context was associated with increased activation of the precuneus (Beyer et al., 2018), a key node in the mentalizing network. This supports the hypothesis of strong interactions between mentalizing and wider cognition. Based on these findings, we developed a cognitive model (Figure 1) of how social context influences sense of agency (Beyer et al., 2017, 2018). This model states that in social contexts, mentalizing interferes with decision-making processes, as the potential actions of other agents must also be considered, thereby reducing sense of agency. This model draws on previous work showing that sense of agency is reduced by dysfluency in action selection (Sidarus et al., 2013, 2017a; Sidarus & Haggard, 2016) and increased cognitive load (Hon et al., 2013; Howard et al., 2016; Wen et al., 2016). Here, we further investigate this framework of how social settings may influence human action processing.

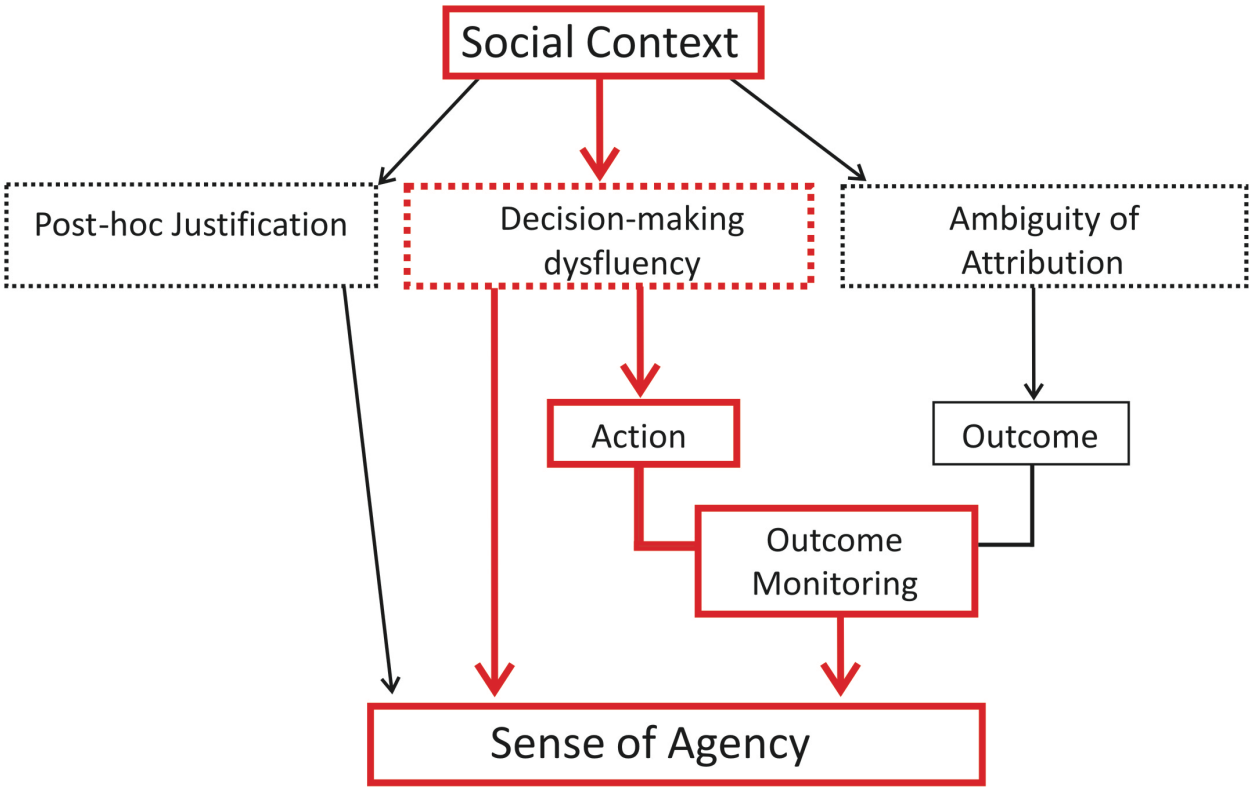


Figure 1: model of social context influences on sense of agency. (from Beyer, Sidarus et al., 2017) The model shows the proposed mechanism behind how the presence of other people can reduce outcome monitoring and sense of agency (shown in red). We propose that in social contexts, mentalizing processes increase dysfluency in the individual's decision-making and action planning process. This dysfluency leads to a subjective loss of control over the outcomes of the individual's own actions. Importantly, we have previously shown that this process is independent of post-hoc reinterpretation or justification of action and outcomes, and of ambiguity about the author of a given event (shown in dashed black lines).

To test the modulation of sense of agency in social and non-social contexts, we designed a task in which participants allegedly interacted with another person, while preserving their objective control over the outcomes of their own actions. In this task, participants made costly actions to avoid a negative event, such as an inflating balloon bursting, as shown in figure 2. In order to mimic the payoff structure of classical bystander scenarios, in which actions such as helping are effortful but necessary, we designed actions to be costly (result in the loss of monetary points),

but not acting – and letting the balloon burst – was even more costly. Importantly, participants had some control over the outcomes of their actions, as they lost fewer points, on average, the later they stopped the balloon. Yet, there was also risk involved in the decision, as the balloon could inflate at different rates across the trials, and could suddenly speed up during the trial.

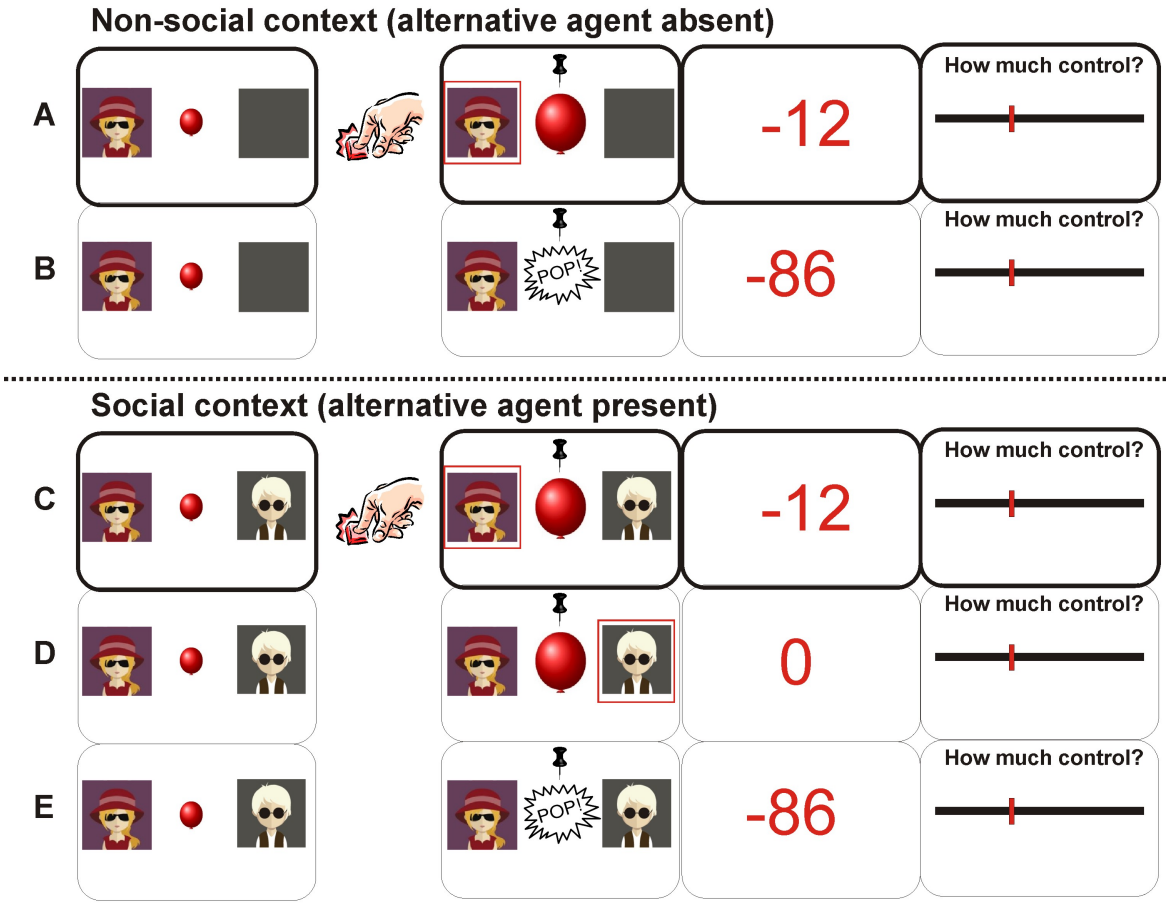


Figure 2: task outline to study social context effects on sense of agency in Experiment 1. Figure shows the different conditions for the task, similarly to previous studies Co-player absent context: participant successfully stops the balloon and loses the respective number of points (**A**); balloon pops, participant loses larger number of points (**B**). Co-player present condition: participant successfully stops the balloon and loses the respective number of points (**C**); co-player stops the balloon, participant loses 0 points (**D**); balloon pops, participant loses larger number of points (**E**). Analyses focused on trial types **A** and **C**.

As shown in figure 2, in some trials, participants played alone, and should decide *when* to act to stop the balloon inflating before it burst, weighing the potential risk costs and against the benefits of acting later. In other trials, participants were told that they were playing with another person, represented on the screen as a second avatar. In those trials, if the co-player acted first to stop the balloon, the participant no longer needed to act and hence would not lose any points. However, if neither player acted, both participants lost a large number of points. Crucially, immediate action feedback – highlighting the avatar of the actor and the stopped balloon – eliminated ambiguity as to who was the author of a given outcome. Nevertheless, when the other player was present, participants' behaviour changed, as they tended to act later to stop the balloon, reported a reduced sense of agency over the outcomes of *their own* actions, and showed reduced outcome monitoring at the neural level (Beyer et al., 2017).

Importantly, our cognitive model of the impact of social context on sense of agency (Beyer et al., 2017, 2018) generates clear, testable hypotheses, which had remained untested and are addressed in the current study. Specifically, if sense of agency is reduced in social contexts due to mentalizing processes interfering with decision-making, then this effect should:

1. Depend on the social nature of the task, wherein the possible behaviour of other agents will be actively considered during decision-making. A non-social context that merely increases uncertainty about upcoming events should not have the same effect.
2. Be independent of outcome valence. Our model assumes that reduced sense of agency is the result of cognitive processes during action selection, rather than of post-hoc evaluation of action outcomes

The current experiments are therefore designed to directly test these hypotheses, to exclude key alternative explanations, while also testing the replicability and generalizability of our previous findings.

Most importantly, our previous studies lacked a non-social control, so the only influence on participants' decisions was a social agent. This meant that social modulation of sense of agency could not be distinguished from a general effect of uncertainty on sense of agency, or a more general change in the perceived risk in the trial, since the social context offered the possibility that not acting could result in a good outcome (i.e. as the balloon could be stopped by the co-player). To address this, the first experiment involves two setups that are identical in terms of the events that participants experience, but differ in their instructions. Namely, one group of participants receive instructions that any external influence on the task is caused by another person. The other group is instructed that any influence is caused by a faulty mechanical device – an "old" balloon pump that can malfunction and stop inflating the balloon. Playing with another person is expected to lead participants to mentalize about the co-player's behaviour, trying to understand and predict when the co-player will act, and incorporating such predictions in their decision-making, in addition to the risk calculations. In contrast, while the faulty pump condition still introduces uncertainty about upcoming events, and could potentially alter the risk calculations, it is not expected to engage additional cognitive processes for modelling and predicting when the pump will fail to inflate the balloon. This allows for a direct test of the influence of social cognition on sense of agency.

While the above setup tests the most important alternative explanation for our previous findings, still another potential influence remains in the tasks used previously. So far, our studies only

involved negative action outcomes, thus we could not exclude the possibility that there was something specific about negative outcomes in social contexts. Generally, participants may be motivated to reduce their personal sense of agency for negative events, in line with the concept of self-serving bias (Bandura, 2002). Yet, even in the presence of a self-serving bias, one could hypothesise different patterns of interaction between social context and outcome value, depicted in Figure 3, that carry different implications for the role of self-serving bias in understanding diffusion of responsibility. Here, outcome value is considered in a relative sense, represented by a Z-score, where 0 represents average outcomes, and more positive vs. negative values represent increasingly better vs. worse than average outcomes, respectively. Classically, it has been assumed that the diffusion of responsibility effect is specifically tied to a self-serving bias, as the presence of another agent would offer an opportunity to strategically displace responsibility, away from the self and towards the other, for undesirable outcomes. Within the context of our task, this hypothesis would predict that agency ratings should be especially reduced in the social, relative to non-social, context for worse outcomes – as depicted under H1 (figure 3). In contrast, our previous studies have shown that participants demonstrated a *general* self-serving bias, giving gradually lower agency ratings with increasingly undesirable (more negative) outcomes (Beyer et al., 2017, 2018), but this effect was the same across social and non-social contexts – as depicted under H2. This suggests that diffusion of responsibility is an independent effect that cannot be explained by a self-serving bias. Finally, one could hypothesise a third pattern of results, H3, wherein the reduction in agency ratings due to a social context would only be evident for more desirable outcomes. In such a scenario, particularly low agency ratings for relatively bad outcomes might result in a floor effect, obscuring the influence of social context. Importantly, results resembling those of either H2 or H3 would show that diffusion of responsibility could not be explained *through* a self-serving bias. Our previous work already supported H2. Yet, it

remains possible that these results were due to actions always having a (more or less) negative outcome, thus creating a situation in which displacing responsibility might be seen as favourable. Therefore, in a second experiment, we tested whether the presence of another agent reduces sense of agency similarly for overall positive vs. overall negative action outcomes.

We discuss the implications of our findings for common practices of education and for our understanding of social development.

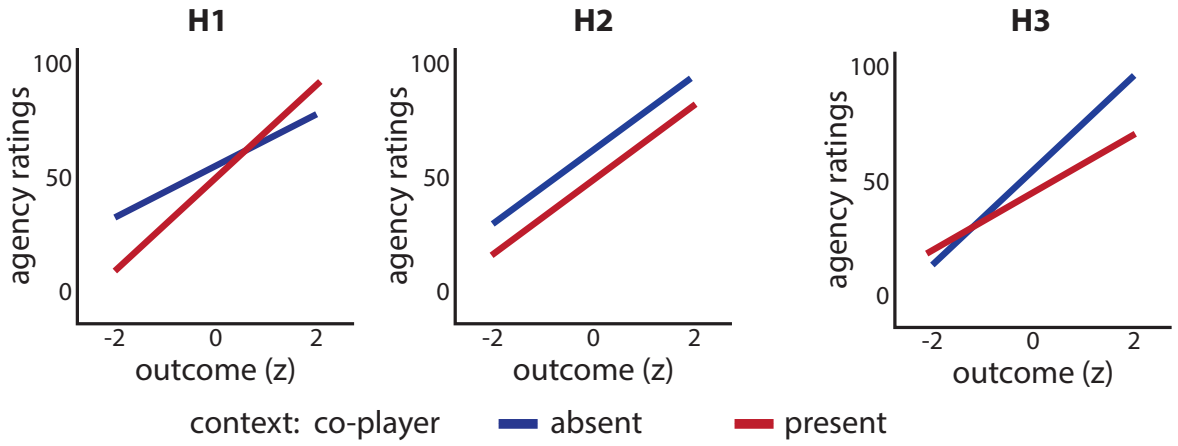


Figure 3: Hypothetical interactions between self-serving bias and diffusion of responsibility. Across the 3 panels, there is an overall self-serving bias, with agency ratings gradually reducing with increasingly less desirable outcomes but each panel carries different implications. Outcome value is here *standardised* (Z-scored), ranging from better than average outcome values, i.e. positive Z values, to average outcomes (0), towards worse than average outcomes, i.e. increasingly negative Z scores. H1: diffusion of responsibility (i.e. lower agency ratings in social, than non-social, context) is due to a self-serving bias, as evidenced by a strategic displacement of agency with more undesirable outcomes. H2: diffusion of responsibility is independent from a self-serving bias. H3: diffusion of responsibility cannot be explained by a self-serving bias, but can be overshadowed by it.

Experiment 1

If people feel less in control in social action contexts because mentalizing processes interfere with decision-making, then this effect should be specific for social influences. However, if mere

uncertainty prior to the action or post-hoc counterfactual thinking leads to the subjective loss of agency, then this should also be observed for non-social sources of alternative trial outcomes. We compared the agency-reducing effect of the presence of an alternative agent between two task settings (figures 2 & 3). Both setups were identical in all aspects, except that the alternative agent was introduced either as a human co-player, or as a non-intentional and non-social mechanical device.

Methods

All measures, manipulations and exclusion of data for the experiments reported here are explained in the manuscript.

Sample size, participants & procedure

For both experiments, we based the experimental methods on previously established findings. The task we used has been shown to result in reliable, replicable within-subject effect of context (i.e. alternative agent absent vs. present; Beyer et al., 2017, 2018). Sample size was determined *a priori* based on previous studies, aiming for N=24 per group, and constrained by participant availability. We planned to test the main effects of interest on agency ratings using multilevel regression models, given their greater sensitivity and reliability relative to standard statistical tests (e.g. ANOVAs) that do not simultaneously model variability in effects across and within participants (Gelman & Hill, 2006; McElreath, 2015). Unfortunately, it remains difficult to perform classic power calculations for multilevel regression models, due to the heterogeneous sources of variance that must be taken into account (McElreath, 2015; Westfall et al., 2014). Therefore, we opted to analyse agency ratings using a Bayesian approach to multilevel regression. Bayesian methods thus allow us to assess the strength of evidence in our data for the effects of interest, given our sample size.

48 healthy volunteers (9 male; age 18-31, mean age = 23; 4 left-handed) were recruited for experiment 1. 24 participants (3 male) performed the task in the social condition, 24 (6 male) performed the task in the non-social condition. No participants were excluded from data analysis. For the social version, participants were invited into the lab in pairs, received instructions together and were told that they would be playing together in the experiment. They were then brought into separate computer cubicles to perform the task. For the non-social version, participants were also recruited in pairs, but were not told they would be playing together. In case one participant failed to attend, the other was assigned to the non-social condition and tested alone (n=9). After the task, participants filled out a post-experimental questionnaire, were fully debriefed and paid £7.50 per hour for their participation, plus a bonus based on their task performance. All participants gave written informed consent and the study was approved by the local ethics committee.

Task

The task was similar to that used in (Beyer et al., 2018) and modelled after the balloon analogue risk task (Lejuez et al., 2002). In each trial, participants saw a small balloon in the centre of the computer screen, which inflated at constant speed. The image of a pin was presented above the balloon, such that the balloon would pop when it touched the pin. The balloon would inflate at variable speed and speed up unpredictably at some point of a given trial, in order to make it risky to wait until the maximum size possible. At any time, participants could stop the balloon by pressing the space bar on a standard keyboard.

In the social version (figure 1), an avatar marked the presence or absence of the alternative agent. To the left of the balloon, the participant saw an avatar representing themselves. To the right of

the balloon, the participant saw either a coloured rectangle (in non-social trials), or another avatar representing their alleged co-player (in social trials). In social trials, the co-player could sometimes stop the balloon before, and thus instead of, the participant. In each trial, the avatar belonging to the player who stopped the balloon was marked by a red rectangle as soon as a response was made.

In the non-social version (figure 4), participants saw the image of an air pump that was coloured either green or blue. Participants were instructed that the green pump was new, and the blue pump was old. The green pump would always inflate the balloon until it popped, unless the participant acted. The blue pump might, on some trials, break down before the balloon was fully inflated, in which case the participant would not lose any points.

Critically, the social "co-player" and the non-social "faulty pump" were programmed in the same way: the alternative agent would only act if the participant had acted on the majority of social/old pump trials and for a maximum of 3 trials per block. The only difference between task versions was that the pump was introduced as a non-social agent, thus not encouraging the engagement of mentalizing processes.

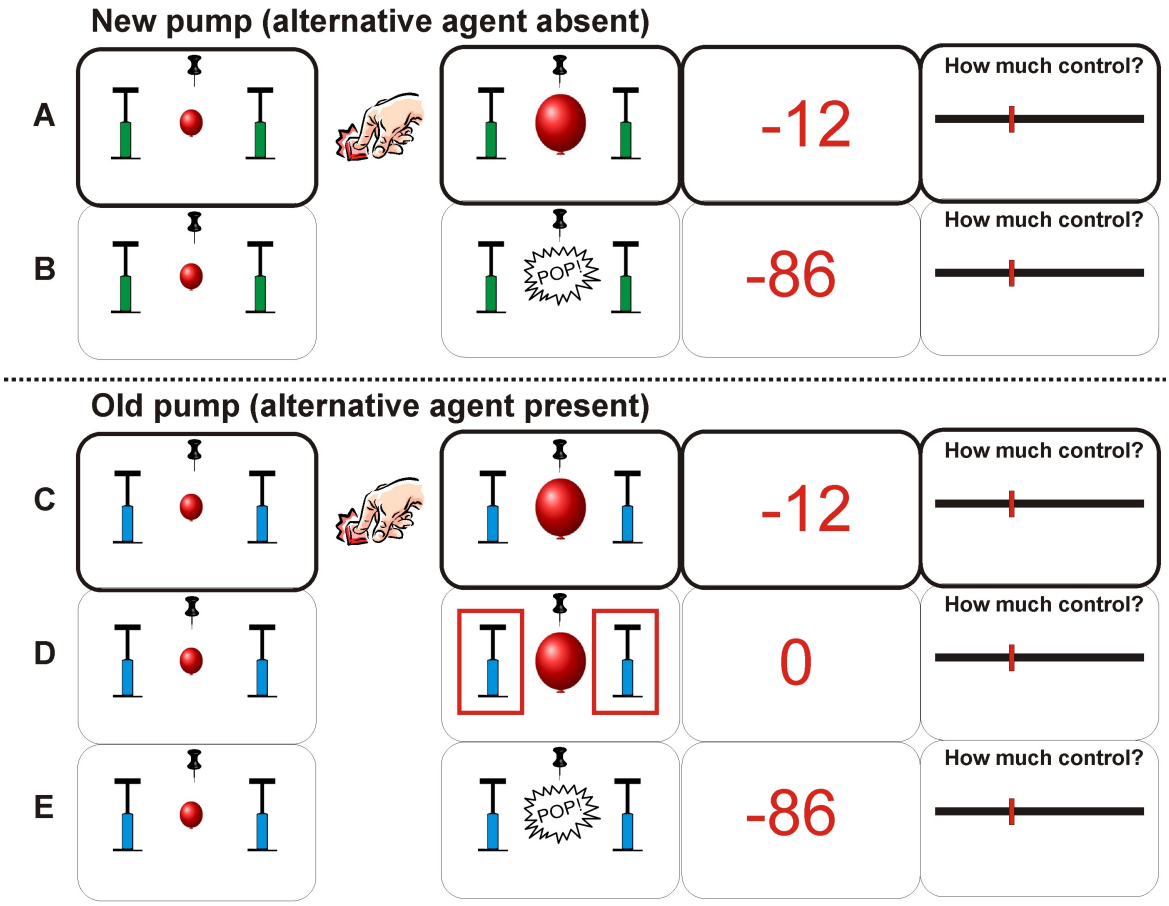


Figure 4: task outline for non-social frame in experiment 1. Figure shows the different conditions for the non-social task version. Within-subject conditions and outcomes were identical to the social task version shown in figure 2.

The payoff structure was as follows: if the balloon popped, participants lost 80-99 points (and the social group was told that, in social trials, so would their co-player); if they stopped the balloon, they lost 1-60 points; in trials with the alternative agent, if that agent stopped the balloon, participants lost 0 points. The other agent (co-player / old pump) was programmed to stop the balloon with a likelihood of about 70%, if the participant had acted on the majority of social trials, and for a maximum of 3 trials per block. The point at which the co-player acted / the old pump broke down varied between 74-86% of the maximum balloon size.

Participants completed three blocks of 20 trials each with 10 agent absent (co-player absent / new pump) and 10 agent present (co-player present / old pump) trials per block, randomized on a trial-wise basis.

After the last block, participants in the social group were given the following questions, answering on visual analogue scales: ‘How fair was your co-player’ (scale labelled as ‘very unfair’ / ‘very fair’); 'When you played together with your co-player, in what percentage of trials did the balloon pop?' (0% / 100%); 'When you played together with your co-player, in what percentage of trials did YOU stop the balloon?'; 'When you played alone, in what percentage of trials did you stop the balloon?'; 'When you played with your co-player, did you believe you were really playing with him/her?' (‘Not at all’ / ‘Completely’). Participants in the non-social group were only given questions 2-4, re-phrased in regard to the old/new pump instead of the co-player.

Data analysis

Our analysis focused on agency ratings in trials in which the participant successfully stopped the balloon before it burst, as these trials are comparable between contexts in which the alternative agent (co-player or old pump) was present or absent.

Analyses were performed with Bayesian multilevel linear regression models (a.k.a. mixed-effects models), with the *brms* package (Bürkner, 2017) in R (R Development Core Team, 2008), which uses Hamiltonian Monte Carlo to sample from the posterior distribution over parameter values, by means of the Stan programming language (Carpenter et al., 2017). We report the posterior means (*b*) of the estimated parameters at the population-level (fixed effects), and their associated 95% credible intervals (CI; the central 95% of values in the respective marginal posterior distribution, indicating the uncertainty around the estimate). We entered trial-wise agency ratings as the dependent variable, modelled by group (social = .5 vs. non-social = -.5) as a between-

subject predictor, with alternative agent context (absent = .5 vs. present = -.5) and outcome value (Z-scored within participant; (Gelman, 2008) as within-subject predictors. The within subject predictors were included as variable effects nested within participants (i.e. random intercepts and slopes model). In a previous study using this paradigm (Beyer, Sidarus et al 2017), we consistently found regression slopes of less than 5 points. Therefore, we specified the prior for the population-level effects a b ~ Normal(0, 5) – that is, Normally distributed with a mean of 0 and standard deviation of 5. This reflects that we are ~95% certain that regression slopes will be within the interval [-10, +10]. We set a Uniform(0, 100) prior on the intercept parameter, covering the range of the scale. We calculated Bayes Factors (BF) for each regression term using the Savage-Dickey density ratio (Wagenmakers et al., 2010). As appropriate, we report effects in favour of the null hypothesis (BF_{01}), or in favour of the alternative hypothesis ($BF_{10} = 1/BF_{01}$, and following (Lee & Wagenmakers, 2014), we describe the strength of evidence as anecdotal ($1 < BF < 3$), moderate ($3 < BF < 10$), strong ($10 < BF < 30$) and very strong ($30 < BF$).

Results

Influence of social context on task performance

Comparing task performance between task versions showed, most importantly, no difference between social (avatar) and non-social (pump) agent groups in the number of trials in which the alternative agent acted (M = 7.6 / 7.5; SD = 1.6 / 1.6; $t_{46} = 0.4$, p = .656; d=.06; Figure 5A). Thus, participants in the social and non-social versions experienced the same level of external influence and, in principle, could have formed similar expectations about the probability of the balloon stopping 'on its own'.

Considering the number of trials in which the participant *did* act, a group by context mixed ANOVA showed significant main effects of group ($F_{1,46} = 8.0$; $p = .007$, $\eta_p^2 = .15$), context ($F_{1,46} = 236.3$; $p < .001$, $\eta_p^2 = .84$), and a significant interaction ($F_{1,46} = 5.6$; $p = .023$, $\eta_p^2 = .11$). Post-hoc tests revealed that, when the alternative agent was present, participants in the social task frame acted less frequently than participants in the non-social frame (M = 16.3 / 19.2; SD = 3.3 / 3.0; $t_{46} = -3.2$; $p = .002$; $d = .92$), while there was no difference between groups when the alternative agent was absent (M = 24.6 / 25.3; SD = 2.5 / 1.9; $t_{46} = -1.0$; $p = .304$; $d = .32$; Figure 5B). While, as is to be expected, both groups acted less often when the balloon could be stopped by the alternative agent (paired t-test for agent present vs. absent, social frame: $t_{23} = 11.7$, $p < .001$; $d = 2.78$; non-social frame: $t_{23} = 10.0$, $p < .001$; $d = 2.32$), this effect was stronger if participants thought they were playing with another person, than if they were playing with a faulty pump. Thus, even though they had the same experience of external influence on stopping the balloon, participants who believed the alternative agent in that condition to be another person relied more on the other agent to act, relative to participants who did not believe that another person was involved. Since both groups had the same number of trials in which the alternative agent acted, acting less often in the agent present condition for the social frame group resulted in a larger number of balloon bursts trials, and hence a slightly inferior task performance, with a lower gain on average (points gained in the social vs. non-social groups: M = 46.6 / 70.6; SD = 33.9 / 21.1; $t_{46} = 2.9$; $p = .005$; $d = .85$).

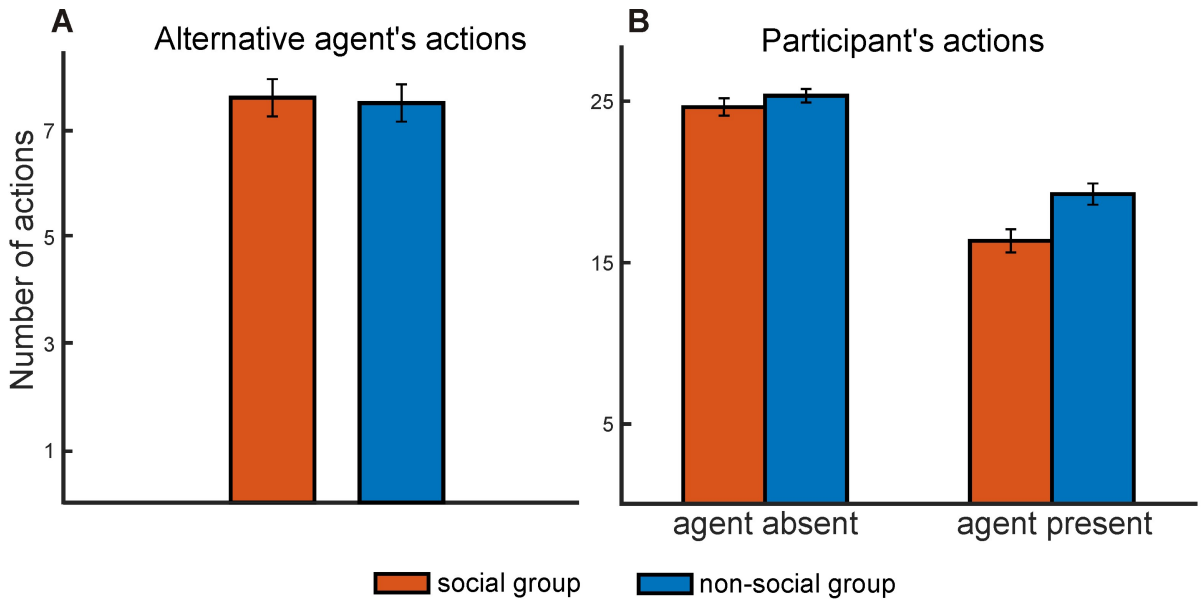


Figure 5: task performance. Panel A shows the mean number of "actions" by the alternative agent, i.e. when co-player acts (social group), or old pump breaks down (non-social group). Panel B shows the mean number of successful actions by the participant in both experimental groups, as a function of the context (agent absent vs. present).

We analysed response times (RTs) with a group (social and non-social groups) x context (agent absent vs. present) mixed ANOVA. This revealed no significant main effect of group ($F_{1,46} = 0.9$; $p = .358$, $\eta_p^2 = .02$) or context ($F_{1,46} = 1.9$; $p = .197$, $\eta_p^2 = .04$), nor a significant interaction ($F_{1,46} = 1.2$; $p = .285$, $\eta_p^2 = .03$; agent absent vs. present for social group: M = 6.35 / 6.33; SD = .22 / .30; agent absent vs. present for non-social group: M = 6.33 / 6.23; SD = .21 / .29). The absence of any effect on RTs in this experiment suggests that changes in its design and the way the behaviour of the alternative agent was programmed, relative to our previous study (Beyer et al., 2017), may have reduced the variance in RTs. Nonetheless, the increased number of balloon bursts in the presence of the *social* agent clearly demonstrates that participants tended to wait for the other player to act.

Influence of social context on sense of agency

Our analyses focused on trials in which the participant stopped the balloon. For these trials, event sequences and action-outcome contingencies were identical in the alternative agent absent vs. present contexts. The Bayesian multilevel regression model of agency ratings (figure 6) showed very strong evidence for a main effect of outcome value (b = 8.95, 95% CI = [5.55, 12.12], BF_{10} > 4×10^4). Importantly, there was moderate evidence for a group × context × outcome interaction (b = 6.01, 95% CI = [0.73, 11.26], BF_{10} = 6.04; figure 6; full statistics in table 1), suggesting that the group manipulation altered the way in which context and outcomes influenced agency ratings.

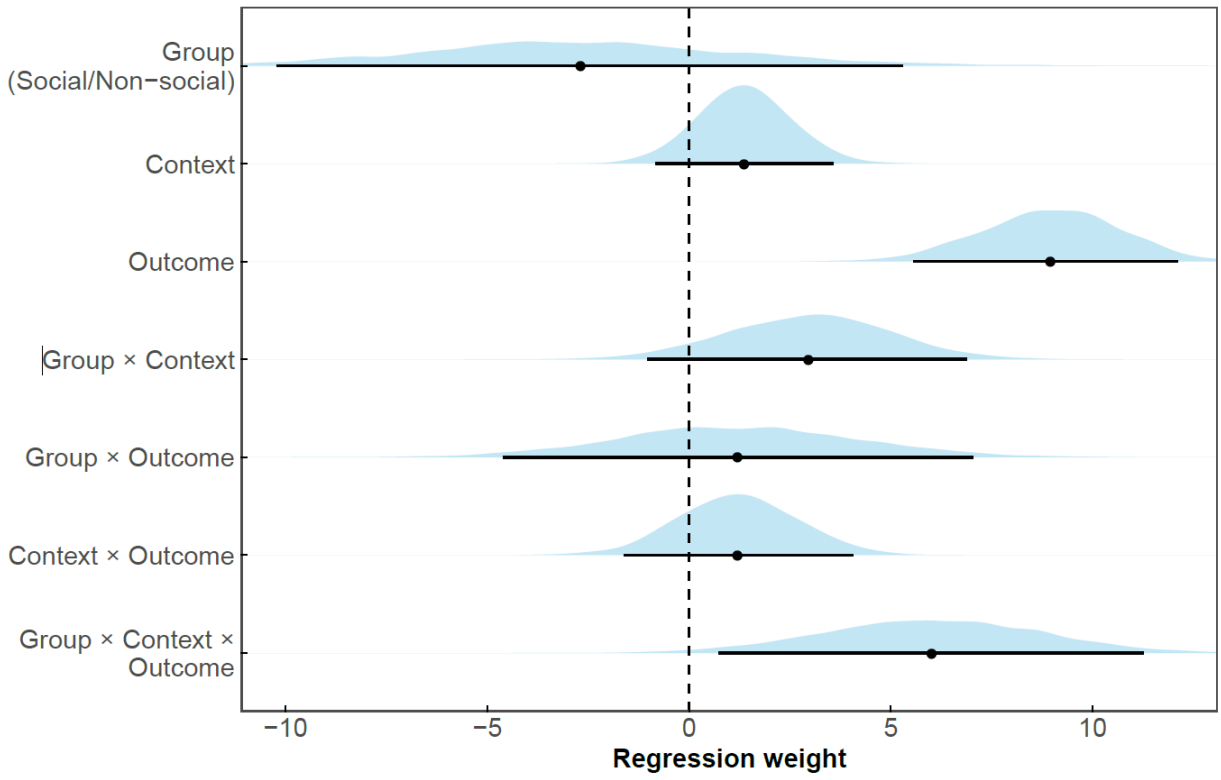


Figure 6: Influences on sense of agency in experiment 1. Density plots of the posterior fixed effects estimates from the Bayesian multilevel model. Points show posterior means, and horizontal lines are 95% Credible Intervals. 'Group' refers to the social (avatar) vs. non-social (pump) factor. 'Context' refers to the presence or absence of the alternative agent (i.e. co-player present/absent, pump old/new).

Table 1: Test statistics for experiment 1. Estimated fixed effect parameters from the Bayesian multilevel model. Columns show the posterior mean estimate, standard error, lower and upper bounds of the 95% Credible Interval, and Bayes Factors in favour of the null (BF_{01}) and alternative (BF_{10}) hypotheses. Group: Social vs. Non-social, Context: presence vs. absence of the alternative agent (i.e. co-player present/absent, pump old/new).

Parameter	Estimate	SE	2.5%	97.5%	BF01	BF10
Intercept	61.32	2.85	55.68	67.00	-	-
Group	-2.69	3.93	-10.22	5.28	0.87	1.15
Context	1.36	1.12	-0.84	3.57	2.04	0.49
Outcome	8.95	1.70	5.55	12.12	$<2.5e\text{-}^4$	$>4\ e^4$
Group x Context	2.96	2.02	-1.04	6.88	0.85	1.18
Group x Outcome	1.19	2.98	-4.60	7.03	1.57	0.64
Context x Outcome	1.20	1.47	-1.62	4.06	2.38	0.42
Group x Context x Outcome	6.01	2.71	0.73	11.26	0.17	6.04
Social Group:						
Context	2.84	1.54	-0.23	5.83	0.61	1.63
Outcome	9.55	2.24	5.15	13.99	$< 2.5\times10^{-4}$	$> 4\times10^3$
Context x Outcome	4.20	2.09	0.21	8.32	0.34	2.97
Non-Social Group:						
Context	-0.12	1.48	-3.07	2.85	3.83	0.26
Outcome	8.35	2.28	3.67	12.71	< 0.01	291.42
Context x Outcome	-1.80	1.91	-5.46	2.02	1.72	0.58

To investigate the three-way interaction, we used our model to estimate the size of the context by outcome interaction within each group (Figure 7). In the social group, we found a context by outcome interaction (b = 4.20, 95% CI = [0.21, 8.32]), with anecdotal evidence for the alternative hypothesis (BF10 = 2.97). In the social group, agency ratings were increasingly greater in the agent-absent context compared to the agent-present context (in which the alleged co-player could have acted) with better outcomes. This interaction resulted in anecdotal evidence for a main effect of context (b = 2.84, 95% CI = [-0.23, 5.83]; BF10 = 1.63), for average outcomes. That is,

the previously observed effect of a reduction in agency ratings in social contexts was here largely restricted to good outcomes, likely due to bad outcomes already leading to a robust reduction in agency ratings, thus overshadowing the context effects.

In contrast, the non-social group showed no robust context by outcome interaction (b = -1.80, 95% CI = [-5.46, 2.02]), with anecdotal evidence *for* the null hypothesis (BF01 = 1.72), nor a main effect of context (b = -0.12, 95% CI = [-3.07, 2.85]), with moderate evidence *for* the null hypothesis (BF01 = 3. 38). Thus, in contrast to the social group, and to our previous findings, the presence or absence of another possible cause for stopping the balloon, i.e. the old vs. new pump, did not robustly affect agency ratings.

Consistent with the large main effect of outcome value in the full model, both groups showed very strong evidence for a main effect of outcome (see table 1), with better outcomes linked to higher agency ratings.

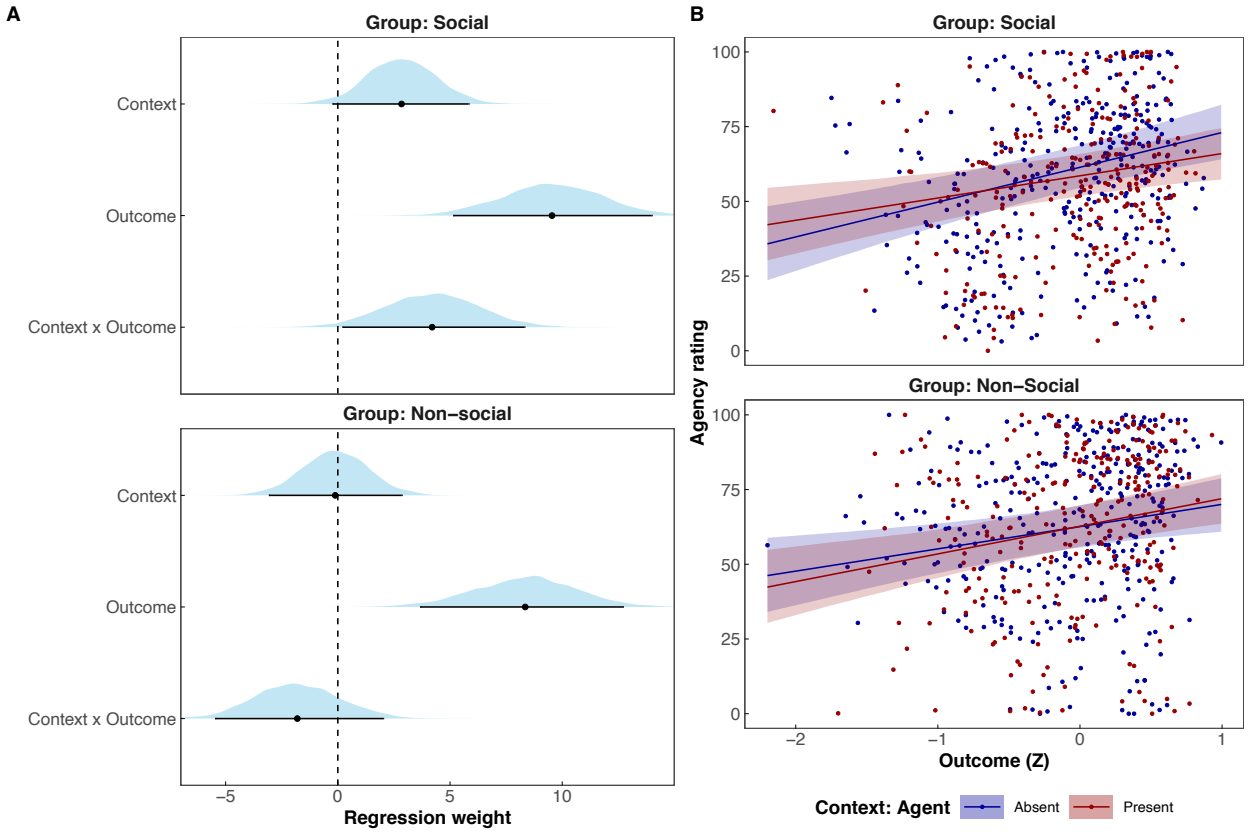


Figure 7: results for separate analysis of social and non-social groups. Panel A shows smoothed density plots of the posterior distributions of the estimated parameters for the effects of context and outcome estimated for the social and non-social group separately. Points show posterior means, and horizontal lines are 95% Credible Intervals. Panel B displays the mean agency ratings (dots) and fitted values from the model (regression line, and shaded 95% Credible Intervals) for the context (alternative agent present vs. absent) by outcome value interactions for each group. Note that more positive outcome values (Z) reflect smaller losses, and more negative values reflect larger losses.

Manipulation checks

At the end of the experiment, participants in the social task group were asked to rate the fairness of their co-player, and whether they had believed they were interacting with the other player, on scales from 0-100%. Participants rated their co-player as moderately fair (M = 47.6%; SD = 22.7) and showed a moderate level of belief in the cover story (M = 54.8%; SD = 35.1). An average rating of >50% indicates that participants were moderately convinced that they were interacting with the other participant. It should be noted that this rating was collected at the very end of the

task, and being given this question itself would likely arouse suspicion. Neither rating was correlated with the effect of social context on sense of agency (fairness: $r = .12$, $p = .59$; belief in cover story: $r = -.06$, $p = .77$). Given this lack of correlation, together with the demand characteristics involved in such debriefing questionnaires, which highlight the possibility of having been deceived, and our use of mixed effects models, which are robust to outliers, we decided to not exclude any participants. These questions were not given to the non-social task group, since there was no alleged other person involved. Including belief ratings a separate predictor in the model of agency ratings showed no main effect of deception, nor any robust interactions (see Supplementary Analysis).

In both conditions, we assessed participants' perception of how many times they acted in either condition. Participants were asked on what percentage of trials they stopped the balloon in social trials / when playing with the old pump. This did not differ between conditions ($M_{social} = 65.2$; $SD_{social} = 14.4$; $M_{non\text{-}social} = 65.7$; $SD_{non\text{-}social} = 18.4$; $t_{46} = -0.1$; $p = .911$). They were also asked on what percentage of social / old pump trials the balloon burst, with participants in the social condition reporting a greater percentage of bursts than participants in the non-social condition ($M_{social} = 38.5$; $SD_{social} = 18.0$; $M_{non\text{-}social} = 27.6$; $SD_{non\text{-}social} = 19.3$; $t_{46} = 2.0$; $p = .05$). For non-social trials / playing with the new pump, there was no difference between groups in the estimated number of times participants stopped the balloon ($M_{social} = 77.9$; $SD_{social} = 15.3$; $M_{non\text{-}social} = 77.5$; $SD_{non\text{-}social} = 19.4$; $t_{46} = .1$; $p = .943$). This demonstrates that participant's impressions of the balloon bursting were largely in line with their actual experience, as the social group experienced more bursts, as presumably they waited for the other agent to act; unlike the non-social group.

Interim discussion

The results of this experiment show that the reduction in sense of agency due to the presence of another potential agent occurs only when that agent is assumed to be a person (i.e. social agent), and not when it is assumed to be a mere mechanism. When a non-intentional, non-social agent could interfere with the balloon inflation in addition to the participant, no reduction in sense of agency was observed for trials in which the participant successfully acted. Participants behaved differently towards social agents, relying more on them than on a non-social agent to intervene in response to increasing risk, and to act before the balloon exploded. These findings show that social cognition is indeed a crucial factor in these contextual effects on sense of agency.

Alternative explanations for reduced sense of agency in the presence of an alternative agent could have been a shift in subjective outcome value when a no-loss option was possible. Thus, due to counterfactual thinking ('I could have lost no points'), a small negative outcome could be perceived as worse than when the no-loss option was not available (in the agent present vs. absent conditions). Further, increased uncertainty of trial outcomes prior to the action, or prior experience of non-control (i.e. the balloon stopping 'on its own'), could become associated with the task condition, thus lowering the overall sense of agency. Crucially, these explanations would have predicted the same effect for the non-social agent, i.e. the old and faulty pump. As the only difference between the two groups was the social vs. non-social framing of why the balloon might occasionally stop "on its own", these findings strongly suggest that social cognition underlies the agency-reducing effect of the co-player's presence.

One other potential difference between conditions could be that the co-player could be perceived as a capable, somewhat predictable aid in the task, whereas the old pump was clearly labelled as

519 defective and random. However, if this had influenced sense of agency ratings, we would have
520 predicted the opposite effects of those found here, i.e. participants should experience particularly
521 low sense of agency when interacting with an unpredictable faulty device.
522 A further difference between task conditions was the presence of a self-representation in the form
523 of an avatar for the social task group, which was absent for the non-social task group. However,
524 for the social group, the participant's own avatar was present in both task conditions (co-player
525 absent or present). Thus, if the presence of such a self-representation affected sense of agency,
526 this should have resulted in a main effect of group, rather than the observed interaction effect.
527
528 In contrast to our previous studies, in the social group here we found evidence for a context by
529 outcome interaction effect, rather than simply a main effect of context. This was due to a stronger
530 effect of the co-player's presence if the outcome of a given trial was relatively good, i.e. fewer
531 points were lost. The most likely explanation for this interaction is a floor effect in agency ratings
532 when outcomes were particularly bad, as participants already rated their sense of agency as very
533 low, thus not reducing it further due to the co-player's presence. Importantly, the direction of this
534 interaction is in the opposite direction of what would be predicted based on self-serving bias,
535 which would predict a stronger displacement of responsibility to others for particularly bad
536 outcomes.
537
538 However, overall negative outcome valence remains a potential confound in the tasks used so far.
539 Previous accounts of diffusion of responsibility have focused on post-hoc justification due to self-
540 serving bias (Bandura, 2002). This predicts that external attribution of control should occur
541 particularly for undesirable outcomes. None of our previous studies found evidence for a stronger
542 effect of social context on sense of agency with increasingly larger losses (Beyer et al., 2017,

2018; Ciardo et al., 2020). In fact, the only interaction between social context and outcomes observed so far showed the opposite pattern, with a reduced effect of social context on sense of agency for particularly negative outcomes.

However, while the effect of social context does not depend on outcome *value* (Z-scored), it may nevertheless be driven by overall outcome *valence*. Particularly, framing outcomes as generally negative could still motivate participants to assign some responsibility to their co-player in social settings, regardless of loss magnitude. As such, a social task frame may simply afford the displacement of responsibility for negative events. To test this alternative explanation, in the second experiment, we compared social context effects on sense of agency for positive and negative outcomes.

Experiment 2

In this experiment, one group of participants performed a "gain" version of the social task (fig. 8), winning a variable amount of points, while another group performed a "loss" version, losing a variable amount of points, as in previous experiments.

Methods

Participants & procedure

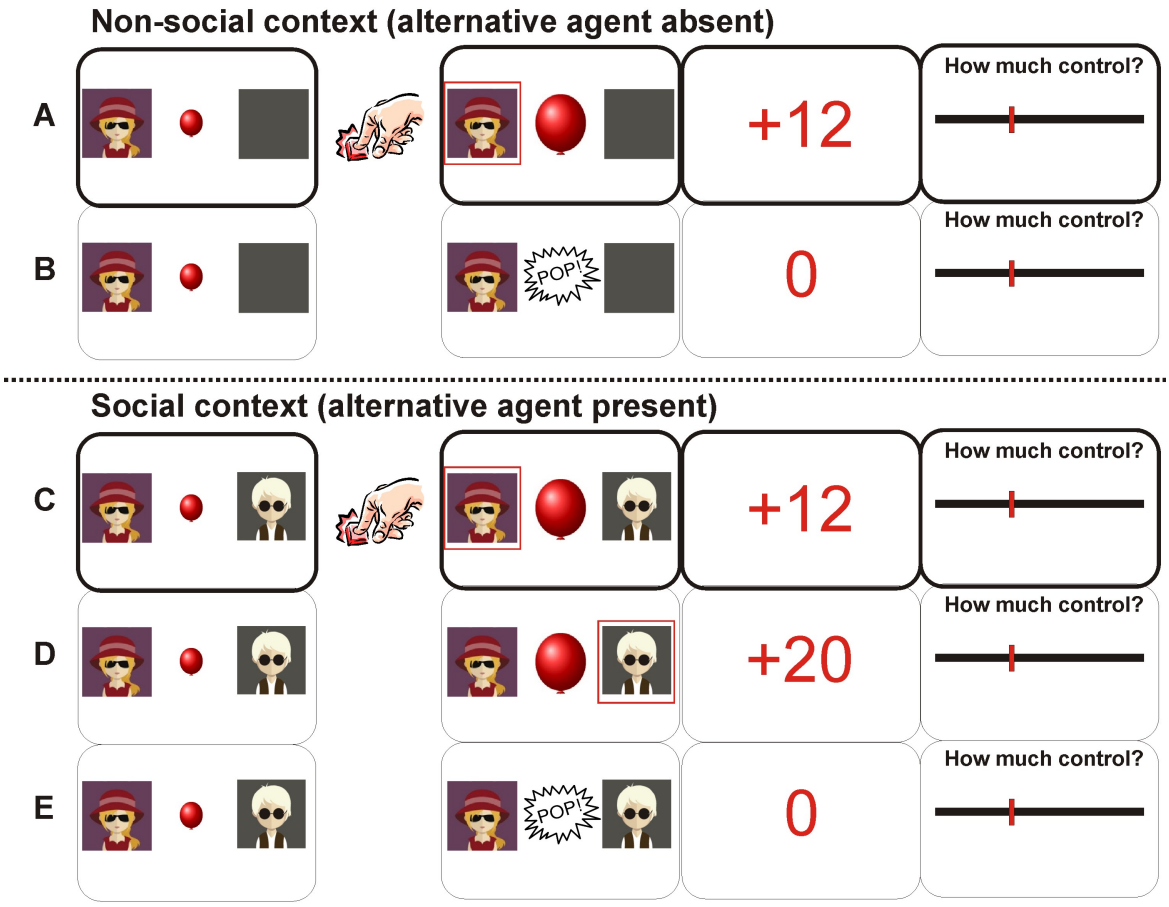
44 healthy female volunteers were recruited for experiment 2. Due to low numbers of male participants being available for testing, only female participants were recruited. 22 participants performed the task in the gain frame, 22 performed the task in the loss frame. One participant in the gain frame was excluded from the analysis due to low trial numbers (only 5 trials in which the

participant successfully stopped the balloon in the social context). Thus, data of 43 participants were included in the analysis (age 19-30, mean age = 23; 2 left-handed).

Participants were invited into the lab in pairs, received instructions together and were told that they would be playing together in the experiment. They were then brought into separate computer cubicles to perform the task. After the task, participants filled out a post-experimental questionnaire, were fully debriefed and paid £7.50 for their participation, plus a bonus based on their task performance. All participants gave written informed consent, and the study was approved by the local ethics committee.

Task

The overall task was similar to that in experiment 1, with the exception that the payoff structure was different, as it needed to be symmetric for the loss and gain version. In the loss frame, the payoff structure was as follows: if the balloon burst, the participant lost 20 points (and was told that in social trials, so would their co-player); if the participant stopped the balloon, they lost 1-20 points depending on the size of the balloon (the bigger the balloon, the fewer points they lost); in social trials, if the co-player stopped the balloon, the participant lost 0 points. In the gain frame, the payoff was as follows: if the balloon burst, the participant earned 0 points; if the participant stopped the balloon, they earned 1-20 points (the bigger the balloon, the more points they earned); in social trials, if the co-player stopped the balloon, the participant earned 20 points. Additionally, there was no pin displayed above the balloon, but the balloon popped at a randomly determined size that varied from trial to trial. At any time, the participant could press the left button on a standard computer mouse to stop the balloon.



589

590 **Figure 8: task outline for experiment 2.** Figure shows the different conditions for the task in the
591 gain frame. Task structure was identical for the loss frame, except for outcome value (which
592 ranged from 0 to -20). In both gain and loss frames, participants obtain the best outcome when
593 the co-player acts, and the worst outcome when the balloon bursts.

594

595 Thus, in both frames, the best outcome was obtained by the co-player's action, the worst if
596 neither player acted, and an outcome in-between these extremes if the participant acted,
597 depending on balloon size. Notably, the overall valence of the outcomes was framed as either
598 something desirable (trying to gain points) or something to be avoided (losing points).
599 At the end of each trial, participants rated how much control they felt they had over the outcome
600 of that trial, on a visual analogue scale ranging from 'no control' to 'complete control'.
601 Participants were instructed that the outcome referred to the number of points they gained or lost
602 on that trial, rather than whether the balloon popped or not.

603

604 The co-player's behaviour was pre-programmed, such that they would only stop the balloon if the
605 participant had stopped the balloon on the majority of social trials of that block (i.e. if the
606 participant had stopped the balloon on at least one social trial more, than the co-player). If this
607 was the case, the co-player stopped the balloon with a likelihood of about 66%.
608 Participants played 4 blocks of 30 trials each. In each block, 15 social and 15 non-social trials
609 were randomly intermixed, resulting in 60 trials per experimental condition.

610

611 *Data analysis*

612 Data analysis was performed as for experiment 1, with Bayesian multilevel linear regression
613 models, with gain and loss frame as a between-subject factor (Gain frame = .5, Loss frame = -.5),
614 with presence of co-player context (absent = .5, present = -.5) and outcome value (standardized to
615 have a standard deviation of 0.5; wherein 0 represents average outcomes, and higher values
616 meaning increasingly more desirable outcomes, i.e. more points gained or fewer points lost) as
617 within-subject predictors. As before, the within subject factors were included as varying effects
618 nested within participants. As in experiment 1, we placed a Normal(0, 5) prior distribution on the
619 fixed effects for all regression parameters, and a Uniform(0, 100) prior on the intercept term.

620

621 Results

622 *Task performance*

623 General task performance did not differ between groups. There was no significant difference
624 across groups in number of trials in which the co-player acted (in the agent present condition;
625 gain vs. loss group: M = 15.62 / 15.73; SD = 3.25 / 2.81; t_{41} = -0.1, p = .908; d = .04; figure 9A),
626 and no significant difference in participants' final earnings (gain vs. loss group: M = 290 / 290;

SD = 24.3 / 21.6; t_{41} = 0.02, p = .983; d = 0). The number of trials in which the participant *did* act was analysed with a group (gain vs. loss frame) by context (agent absent vs. present) mixed ANOVA. This showed no significant effect of group ($F_{1,41}$ < .1, p = .953, η_p^2 < .01), nor a significant interaction between the factors ($F_{1,41}$ = .1, p = .817, η_p^2 < .01; figure 9B). A significant main effect of context ($F_{1,41}$ = 221.8, p < .001, η_p^2 = .84) showed that, across groups, participants acted significantly less often when the alternative agent was present than absent, since the balloon could also be stopped by the co-player (agent absent vs. present for gain group: M = 46.4 / 30.9; SD = 6.7 / 7.2; t_{20} = 9.63; p < .001; d = 2.23; agent absent vs. present for loss group: M = 46.1 / 31.0; SD = 5.8 / 5.9; t_{21} = 11.61; p < .001; d = 2.58).

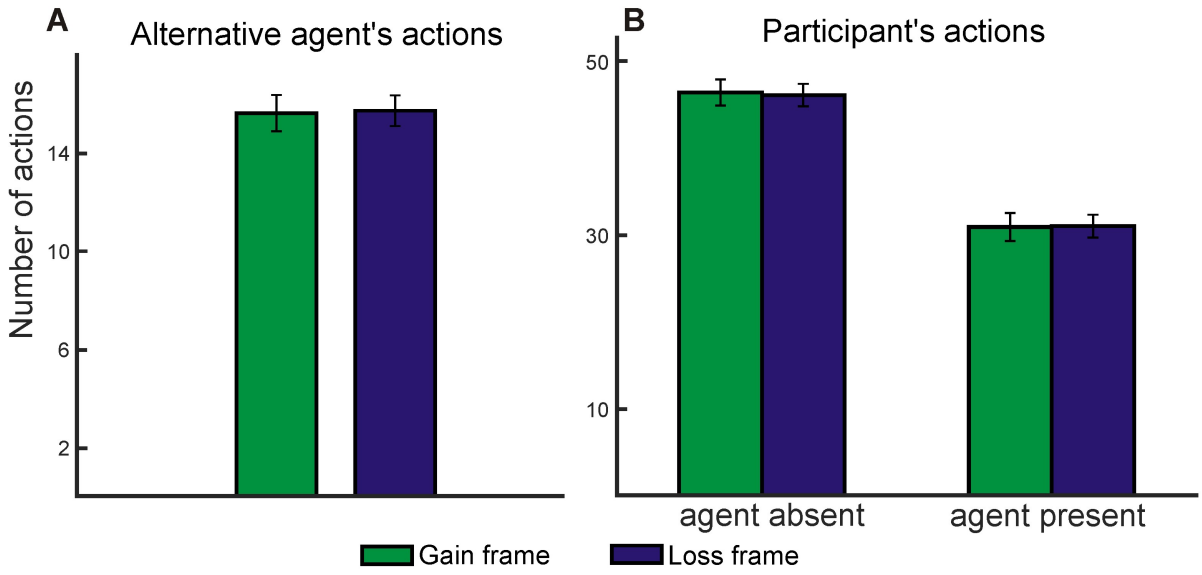


Figure 9: task performance for experiment 2. Figure shows mean number of the alternative agent's actions (co-player acts), as well as mean number of successful actions of the participants in both experimental groups.

Analysis of RTs with the same mixed ANOVA revealed no significant main effect of group ($F_{1,41}$ = 0.1; p = .759, η_p^2 < . 01), nor a significant interaction ($F_{1,41}$ = 1.3; p = .267, η_p^2 = .03). A significant main effect of context ($F_{1,41}$ = 27.4; p < .001, η_p^2 = .40) showed that, across both groups, participants acted significantly later in the agent present than in the agent absent

condition (agent absent vs. present for gain group: M = 6.4 / 6.7; SD = .5 / .4; agent absent vs. present for loss group: M = 6.5 / 6.7; SD = .4 / .3). Consistent with our previous findings (Beyer et al 2017), this suggests that participants tended to wait a bit longer to act when an alternative agent was present, since the best outcome was obtained if the co-player acted instead of them. Importantly, participants' behaviour was equally affected by the co-player across gain and loss groups.

Influence of outcome valence on sense of agency and its modulation by social context

As before, our analyses focused on trials in which the participant stopped the balloon, in which event sequences and action-outcome contingencies were identical for trials with a co-player present vs. absent. The Bayesian multilevel regression model of agency ratings included the predictors group (gain vs. loss frame), context (co-player absent vs. present) and outcome (standardized). This revealed strong evidence for a main effect of context (b = 3.01, 95% CI = [1.09, 4.90], BF_{10} = 18.3), as well as strong evidence for a context × outcome interaction (b = 3.50, 95% CI = [1.32, 5.66], BF_{10} = 24.4, and very strong evidence for a main effect of outcome value (b = 9.73, 95% CI = [6.82, 12.55], $BF_{01} > 4\times10^4$); see figure 9, and full statistics in table 2). Consistent with the social group in Exp. 1 and previous findings (Beyer et al., 2017, 2018), participants felt more in control over better outcomes, and felt less in control in the social context, when a co-player was present, compared to the non-social one, when playing alone. Importantly, as for experiment 1, the interaction between outcome value and social context demonstrates that a self-serving bias, leading to a strategic displacement of agency for undesirable outcomes, cannot explain the reduction in agency ratings in the social context. As figure 10B shows, the difference in agency ratings between social and non-social context increased for better outcomes, and was absent for particularly bad outcomes.

669

670 Crucially, we found anecdotal evidence against an interaction between gain/loss group and
671 context (b = 0.87, 95% CI = [-2.65, 4.29], BF_{01} = 2.55), and anecdotal evidence against a group x
672 context x outcome interaction (b = 0.76, 95% CI = [-3.41, 4. 87], BF_{01} = 2.23). Finally, we found
673 anecdotal evidence against both other effects involving the group term (main effect of group: b =
674 -2.84, 95% CI = [-10.04, 4.34], BF_{01} = 1.08; group x outcome: b = -2.46, 95% CI = [-7.53, 2.71],
675 BF_{01} = 1.26). Together, these findings support our prediction that the previously observed
676 reduction in agency ratings in the presence of intentional agents was not related to the overall
677 context of losing money, as similar effects were observed in the context.

678

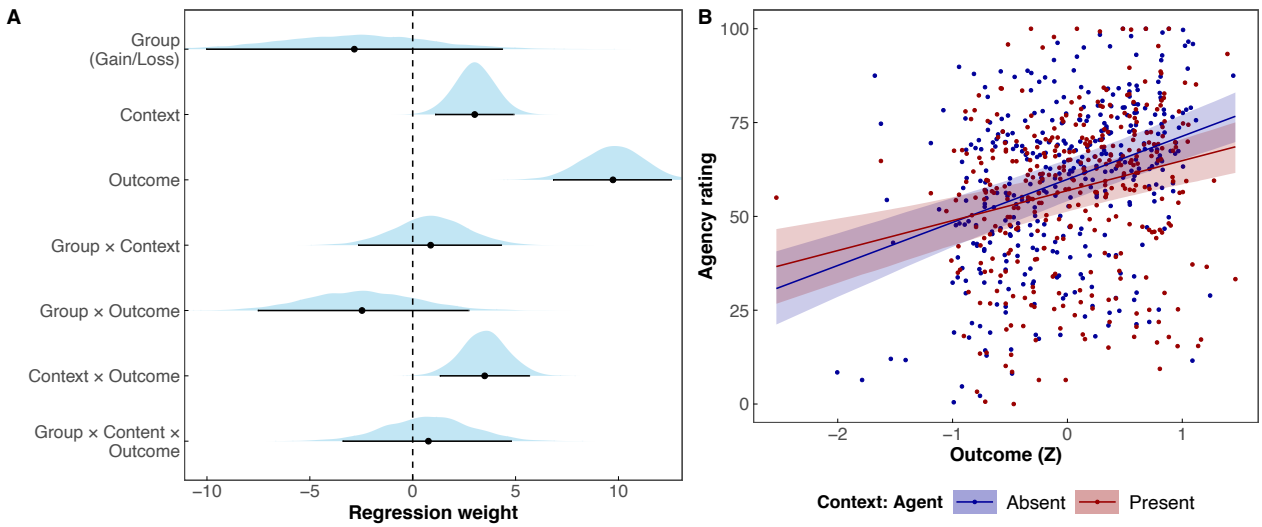
679 **Table 2: Test statistics for experiment 2.** Estimated parameters at the population-level from the
680 Bayesian multilevel model. Estimate is the posterior mean and SE is the posterior standard deviation, with
681 lower and upper bounds of 95% credibility intervals. Group: Gain vs. Loss frame, Context: presence
682 vs. absence of the alternative agent (i.e. co-player present/absent).
683

Parameter	Estimate	SE	2.5%	97.5%	BF01	BF10
Intercept	58.37	2.68	52.66	63.48	-	-
Group	-2.84	3.67	-10.04	4.34	1.08	0.93
Context	3.01	0.96	1.09	4.90	0.05	18.3
Outcome	9.73	1.47	6.82	12.55	$<2.5\times10^{-3}$	$>4\times10^{4}$
Group x Context	0.87	1.75	-2.65	4.29	2.55	0.39
Group x Outcome	-2.46	2.62	-7.53	2.71	1.26	0.80
Context x Outcome	3.50	1.10	1.32	5.66	0.04	24.4
Group x Context x Outcome	0.76	2.09	-3.41	4.78	2.23	0.45

684

685

686



687

688 **Figure 10: Influences on sense of agency for experiment 2. A.** Density plots of the posterior
689 distributions of the estimated parameters at the population-level from the Bayesian multilevel
690 model. Points show posterior means, and horizontal lines are 95% Credible Intervals. 'Group'
691 refers to the gain vs. loss frame. 'Context' refers to the presence or absence of the alternative
692 agent (i.e. co-player present/absent). **B.** Mean agency ratings (dots) and fitted values from the
693 model (regression line, and shaded 95% Credible Intervals) for the context × outcome value
694 interaction effect, collapsed across loss and gain frame groups. Note that more positive outcome
695 values (Z) reflect smaller losses or larger gains (loss/gain group), and more negative values
696 reflect larger losses or lower gains, respectively.

697

698 *Manipulation checks*

699 Ratings of fairness (M = 48.9%; SD = 17.2) and believing the cover story (M = 52.9%; SD =
700 22.1) were similar to experiment 1 and did not differ between win/loss groups (fairness Win vs.
701 Loss, M = 50.6 / 47.2; SD = 17.3 / 17.5; t_{41} = .66; p = .514; d = .20; believe Win vs. Loss, M =
702 49.3 / 56.7; SD = 21.3 / 22.9; t_{41} = 1.11; p = .274; d = .33). Including belief ratings a separate
703 predictor in the model of agency ratings showed no robust evidence for a main effect of
704 deception, nor any interactions (see Supplementary Analysis).

705

706 <u>Interim Discussion</u>

Our findings show that reduced sense of agency in social contexts is not limited to situations in which action outcomes are undesirable, but also occurs for overall positive outcomes. This is in line with the hypothesis that the reduction in sense of agency in social contexts is driven by mentalizing processes, rather than self-serving bias. Across gain and loss frame settings, for relatively average or good outcomes, participants felt less in control over the consequences of their own actions when another potential agent was present. Thus, reduced sense of agency in social context does not depend on a generalised motivation to displace or diffuse responsibility for negative action consequences. In fact, as seen for the social group of Exp 1, the context by outcome interaction showed that the effect of context increased with more positive outcomes.

Discussion

This study tested key predictions derived from our novel model on how social contexts affect an important non-social aspect of human cognition, namely the emergence of a sense of agency. In a first experiment, we showed that social context reduces sense of agency, particularly for good outcomes, but a comparable, non-social, non-intentional influence in the task did not have this effect. In a second study, we showed that the presence of another social agent led participants to feel less in control over the consequences of their actions, regardless of whether those consequences involved overall financial gains or losses. Importantly, in both cases, the alternative agent had no influence on the outcomes of the participant's action.

Our findings replicate our previous studies using similar tasks, while significantly extending our understanding of important phenomena in social psychology. Generally, differences in human behaviour between non-social and social environments are explained with self-serving biases (Shepperd et al., 2008), shyness or social referencing (DiMenichi & Tricomi, 2018), or strategic

731 displacement of responsibility (Bandura, 2002). Moreover, social contexts can objectively reduce
732 control over one's actions and outcomes, and can introduce ambiguity in who caused a given
733 outcome. Perceived control is an important prerequisite for responsibility: one should reasonably
734 assume more responsibility for a controllable event than for a non-controllable one. We show that
735 the presence of others affects the human experience of voluntary action, even when alternative
736 influences as the ones above are experimentally controlled for.
737
738 In reference to the possible relation between a self-serving bias and diffusion of responsibility
739 described in the introduction, we found no evidence to support the hypothesis that the diffusion
740 of responsibility effect is *specifically tied* to a self-serving bias, such that participants
741 *strategically* displace responsibility to others for undesirable outcomes, as exemplified in H1
742 (figure 3). The second experiment showed a similar reduction in agency ratings in the alleged
743 presence of a co-player, relative to playing alone, i.e. diffusion of responsibility, regardless of
744 whether participants aimed to earn points (gain frame) or avoid losing points (loss frame).
745 Turning to how agency ratings were affected by *relatively* more desirable vs. more undesirable
746 outcomes (i.e. within-participants), our findings are consistent with a *general* self-serving bias, as
747 participants report greater control over better outcomes, but that cannot explain the reduced sense
748 of control in social contexts. If anything, the interaction pattern observed here was of a greater
749 effect of social context on the sense of control with relatively better outcomes, consistent with the
750 pattern of H3 (figure 3). Yet, we suggest this pattern is best explained by a floor effect on ratings
751 for the more undesirable outcomes, which would overshadow the social context effect. When
752 considered together with our previous studies (Beyer et al., 2017, 2018; Ciardo et al., 2020)
753 consistenly showing no interactions between outcome value and social context, as depicted in H2
754 (figure 3), we believe the balance of evidence is most consistent with the hypothesis that the

sense of agency is independently influenced by a self-serving bias, reflected in the effect of outcome, and the diffusion of responsibility seen in social contexts.

Further supporting a dissociation between the effect on sense of agency of social context and of outcome value, higher sense of agency for better outcomes was even observed in a completely non-social task setup (when participants interacted with a pump, Exp 1). Moreover, studies using implicit measures of sense of agency in non-social settings (Christensen et al., 2016; Takahata et al., 2012) have shown a consistent pattern of results, suggesting that this effect does not require explicit, reflective processes. The observed effect of outcome on sense of agency is consistent with a general self-serving bias, such that participants accept more control over actions with more desirable consequences. Yet, a second explanation worth noting would be that participants aimed to achieve the best outcome possible, and thus felt most in control when the observed outcome closely matched that intention.

Together, the two experiments presented here provide strong support for our model of social context influences on sense of agency, developed in earlier studies (Beyer et al., 2017, 2018). According to this model, the presence of others increases dysfluency in the decision-making process, by evoking mentalizing processes in addition to task-directed cognition. This dysfluency then decreases sense of agency, in line with studies demonstrating reduced sense of agency with increased decision-making difficulty (Chambon et al., 2014; Sidarus et al., 2017b; Sidarus & Haggard, 2016; Wenke et al., 2010) or increased working memory demands (Hon et al., 2013; Howard et al., 2016; Wen et al., 2016).

We propose that the presence of another human agent is a particularly strong source of dysfluency, due to the complexity of cognitive processes induced by their presence. Recall that, in the first experiment comparing social and non-social agents, participants in both groups experienced the same amount of external influence in the task, that is, the balloon was stopped by the alternative agent (co-player or faulty pump) in the same number of trials. Yet, the presence of another potential agent only influenced sense of agency when the agent was believed to be a social, intentional entity, compared to a non-living, presumably random one. Since the only difference between groups was the framing of the task, differences in the effects of context on sense of agency between groups likely depend on the cognitive processes associated with the two task versions. Given that the key difference was whether or not the task instructions involved another person, mentalizing processes are the most plausible cognitive process to differ between groups, as is supported by our previous MRI study (Beyer et al., 2018). Plausibly, people try to build a model of the other putative social agent's behaviour in order to predict what the other agent will do. Mentalizing about their co-player's potential behaviour, and trying to predict when and why the co-player might act, would thus serve to help the participant try to avoid the cost of acting themselves. In contrast, participants in the non-social condition were less influenced by their previous experience of the faulty pump, and tended to ignore the influence of the pump during decision-making. This may be because participants could not, or did not expect to, form a predictive model of the pump's relevant behaviour. When the potential alternative cause of the balloon stopping was non-social (i.e. the "old pump"), it might seem *a priori* less predictable, hence, participants might not engage resources in trying to understand its behaviour.

In fact, similar effects have recently been found for interactions with a robot (Ciardo et al., 2020), in a task setting that did not involve monetary payoff, further suggesting that the perception of intentionality (as suggested even by an inanimate, but interactive robot) is sufficient to induce a

reduction in sense of agency. Taking these findings together thus supports our account that assuming an intentional stance towards the social agent results in continuous efforts at modelling and predicting their behaviour. Attempting to form this additional predictive model in turn disrupts the participant's own decision-making and sense of agency.

Our interpretation of our findings as supporting a critical role for mentalizing in interfering with decision-making is further supported by the observation that participants' decisions were indeed different in social contexts. Participants relied more on the alternative *social* agent to act, even to their own disadvantage, as it resulted in more trials in which the balloon popped. This suggests that in addition to deciding *when* to stop the balloon on a given trial, in the presence of a social agent, participants may have additionally considered *whether* they should act at all. This decision would depend on their prediction of the co-player's behaviour. The non-social cause of "action" still increased uncertainty about what might happen in each trial, as the balloon might still stop "on its own". However, participants acted more frequently in this condition, experiencing fewer balloon burst. Thus, only social agents led to robust changes in the participants' decision-making processes, by considering the other's behaviour, in turn disrupting their sense of agency. In line with this, inter-individual differences in perspective taking have been related to susceptibility to the bystander effect, with participants higher in perspective taking traits being more strongly affected by the presence of bystanders (Hortensius et al., 2016).

<u>Limitations and future directions</u>

Alternative explanations for our findings should also be considered. Especially when comparing the social vs. non-social task setups, it is possible that these tasks differed in terms of emotional processes, in addition to cognitive effects. For example, participants could have experienced

826 interaction with another person as competitive or provocative. Further, it is possible that a
827 socioeconomic setting, in which one's own losses contribute to a co-player's gain, may affect
828 sense of agency differently than a non-economic setting. However, the structure of the task and
829 instructions were such that it could also be perceived as a collaborative, turn-taking game. While
830 participants have the individual goal of maximising their own payoff, they also have the shared
831 goal of preventing the balloon from bursting. In fact, as the co-player's behaviour was rated as
832 moderately fair, we consider it unlikely that the observed loss of agency in social settings is
833 primarily due to socioeconomic trade-off considerations, or anger.
834 While our core findings are in line with previous studies, the interaction between outcome
835 magnitude and social context effects has not previously been found. We believe floor effects are
836 the most likely reason for the absence of a social context effect in trials with relatively bad
837 outcomes. Nonetheless, it remains possible that deciding to act early could have altered the effect
838 of social context on sense of agency, which could be explored in future studies. In the current
839 task, response times were partially related to outcome magnitude, rendering it difficult to estimate
840 the potentially specific role of response time on the effect of social context on sense of agency.
841 However, the task was designed such that the speed at which the balloon inflated varied both
842 across and within trials, ensuring that was no strict relationship between response time and
843 outcome magnitude. Notably, there was no strong and consistent effect of social context on
844 response times. Therefore, we do not think this is likely to be a significant confound for the
845 effects observed here.
846 Further, we mostly tested female participants here. However, in a previous study with a balanced
847 gender distribution, we found no evidence of gender effects (Beyer et al., 2017).
848

It remains to be tested whether this agency-reducing effect of social context depends on the nature of the interaction. In the present experiment, the interaction was semi-competitive. In situations where participants engage in a fully shared goal (e.g. joint action setups), or in which a clear rule-based strategy is offered (such as prescribed turn-taking), the effect of the other's presence on sense of agency might be absent or even reversed (cf. van der Wel, 2015). Relatedly, future studies could further address the potential role of perceived uncertainty of the alternative agent, as this may have differed between the social and non-social task groups in experiment 1. One possibility is manipulating the predictability of the co-player's behaviour, to assess whether a more random behavioural pattern affects sense of agency differently than a more strategic or predictable one.

Sense of agency is related to a number of perceptual processes (Tsakiris & Haggard, 2005) and outcome monitoring (Bednark & Franz, 2014), and is thus presumed to play a crucial role in voluntary action. Previous research has largely focused on the benefits of social contexts to human cognition (Devaine et al., 2014; Vanlangendonck et al., 2018). This has neglected its potentially disruptive effects under some circumstances, as when social context reduces sense of agency and outcome monitoring (Beyer et al., 2017). Our findings have strong implications for common educational practices: reduced sense of agency in social contexts may likely affect feedback-driven learning, making a case for reduced peer influence on individual learning processes. Moreover, future studies should take into account interpersonal variability in the sensitivity to social cues, to better understand the role of mentalizing processes in learning from social feedback, and consequently on social development.

Conclusions

In the presence of other people, mentalizing processes can interfere with non-social aspects of human cognition. In two experiments, we show that the presence of others reduces sense of agency over gain and loss outcomes, and that this effect is specific to the presence of an intentional, social agent. Our findings suggest that the presence of other people can have fundamental effects on how we perceive our own actions and outcomes. This has important implications for our understanding of human behaviour in social environments. Even without an explicit motivation for self-serving displacement of responsibility, the presence of others can affect our subjective sense of agency. An anticipated lack of control might reduce an individual's motivation to take action in a social situation, while reduced outcome monitoring could be linked to reduced learning from action consequences. Thus, further studies should focus on the effects that a reduced sense of agency in social situations might have on subsequent learning and decision-making.

Open Practices

Data is available in de-identified form on Open Science Framework (https://osf.io/2s7kb/).

Bandura, A. (2002). Selective Moral Disengagement in the Exercise of Moral Agency. *Journal of Moral Education*, *31*(2), 101–119. https://doi.org/10.1080/0305724022014322

Bednark, J. G., & Franz, E. A. (2014). Agency attribution: Event-related potentials and outcome monitoring. *Experimental Brain Research*, *232*(4), 1117–1126. https://doi.org/10.1007/s00221-014-3821-4

Beyer, F., Sidarus, N., Bonicalzi, S., & Haggard, P. (2017). Beyond self-serving bias: Diffusion of responsibility reduces sense of agency and outcome monitoring. *Social Cognitive and Affective Neuroscience*, *12*, 138–145.

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. *ENeuro*, *5*(1), ENEURO.0336-17.2018. https://doi.org/10.1523/ENEURO.0336-17.2018

Blakemore, S. J., Wolpert, D. M., & Frith, C. D. (2002). Abnormalities in the awareness of action. *Trends in Cognitive Sciences*, *6*(6), 237–242. https://doi.org/10.1016/S1364-6613(02)01907-1

Bolt, N. K., Poncelet, E. M., Schultz, B. G., & Loehr, J. D. (2016). Mutual coordination strengthens the sense of joint agency in cooperative joint action. *Consciousness and Cognition*, *46*, 173–187. https://doi.org/10.1016/j.concog.2016.10.001

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1).

Chambon, V., Sidarus, N., & Haggard, P. (2014). From action intentions to action effects: How does the sense of agency come about? *Frontiers in Human Neuroscience*, *8*, 320. https://doi.org/10.3389/fnhum.2014.00320

Christensen, J. F., Yoshie, M., Di Costa, S., & Haggard, P. (2016). Emotional valence, sense of agency and responsibility: A study using intentional binding. *Consciousness and Cognition*, *43*, 1–10.

Ciardo, F., Beyer, F., De Tommaso, D., & Wykowska, A. (2020). Attribution of intentional agency towards robots reduces one's own sense of agency. *Cognition*, *194*, 104109. https://doi.org/10.1016/j.cognition.2019.104109

Devaine, M., Hollard, G., & Daunizeau, J. (2014). The Social Bayesian Brain: Does Mentalizing Make a Difference When We Learn? *PLOS Computational Biology*, *10*(12), e1003992. https://doi.org/10.1371/journal.pcbi.1003992

DiMenichi, B. C., & Tricomi, E. (2018). Increases in brain activity during social competition predict decreases in working memory performance and later recall. *Human Brain Mapping*, *38*(1), 457–471. https://doi.org/10.1002/hbm.23396

Frith, C. D., & Haggard, P. (2018). Volition and the Brain – Revisiting a Classic Experimental Study. *Trends in Neurosciences*, *41*(7), 405–407. https://doi.org/10.1016/j.tins.2018.04.009

924 Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in*
925 *Medicine*, *27*(15), 2865–2873. https://doi.org/10.1002/sim.3107

926 Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*.
927 Cambridge University Press.

928 Hare, B. (2011). From Hominoid to Hominid Mind: What Changed and Why? *Annual Review of*
929 *Anthropology*, *40*(1), 293–309. https://doi.org/10.1146/annurev-anthro-081309-145726

930 Hon, N., Poh, J.-H., & Soon, C.-S. (2013). Preoccupied minds feel less control: Sense of agency is
931 modulated by cognitive load. *Consciousness and Cognition*, *22*(2), 556–561.
932 https://doi.org/10.1016/j.concog.2013.03.004

933 Hortensius, R., Schutter, D. J. L. G., & de Gelder, B. (2016). Personal distress and the influence of
934 bystanders on responding to an emergency. *Cognitive, Affective & Behavioral Neuroscience*, *16*, 672–
935 688. https://doi.org/10.3758/s13415-016-0423-6

936 Howard, E. E., Edwards, S. G., & Bayliss, A. P. (2016). Physical and mental effort disrupts the implicit
937 sense of agency. *Cognition*, *157*(Supplement C), 114–125.
938 https://doi.org/10.1016/j.cognition.2016.08.018

939 Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge
940 University Press.

941 Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., Strong, D. R., & Brown,
942 R. A. (2002). Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART).
943 *Journal of Experimental Psychology: Applied*, *8*(2), 75. https://doi.org/10.1037/1076-898X.8.2.75

944 Li, P., Han, C., Lei, Y., Holroyd, C. B., & Li, H. (2011). Responsibility modulates neural mechanisms of
945 outcome processing: An ERP study: Modulation of outcome processing by responsibility.
946 *Psychophysiology*, *48*(8), 1129–1133. https://doi.org/10.1111/j.1469-8986.2011.01182.x

947 McElreath, R. (2015). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman
948 and Hall/CRC.

949 Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory.
950 *Behavioral and Brain Sciences*, *34*(2), 57–74. https://doi.org/10.1017/S0140525X10000968

951 R Development Core Team. (2008). *R: A language and environment for statistical computing.* R
952 Foundation for Statistical Computing. http://www.R-project.org

953 Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-
954 analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, *42*, 9–34.
955 https://doi.org/10.1016/j.neubiorev.2014.01.009

956 Shepperd, J., Malone, W., & Sweeny, K. (2008). Exploring Causes of the Self-serving Bias. *Social and*
957 *Personality Psychology Compass*, *2*(2), 895–908. https://doi.org/10.1111/j.1751-9004.2008.00078.x

958 Sidarus, N., Chambon, V., & Haggard, P. (2013). Priming of actions increases sense of control over
959 unexpected outcomes. *Consciousness and Cognition*, *22*(4), 1403–1411.
960 https://doi.org/10.1016/j.concog.2013.09.008

961 Sidarus, N., & Haggard, P. (2016). Difficult action decisions reduce the sense of agency: A study using the
962 Eriksen flanker task. *Acta Psychologica*, *166*, 1–11. https://doi.org/10.1016/j.actpsy.2016.03.003

963 Sidarus, N., Vuorre, M., & Haggard, P. (2017a). How Action Selection Influences the Sense of Agency: An
964 ERP study. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2017.02.015

965 Sidarus, N., Vuorre, M., & Haggard, P. (2017b). How action selection influences the sense of agency: An
966 ERP study. *NeuroImage*, *150*, 1–13. https://doi.org/10.1016/j.neuroimage.2017.02.015

967 Synofzik, M., Vosgerau, G., & Voss, M. (2013). The experience of agency: An interplay between
968 prediction and postdiction. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00127

969 Takahata, K., Takahashi, H., Maeda, T., Umeda, S., Suhara, T., Mimura, M., & Kato, M. (2012). It's not my
970 fault: Postdictive modulation of intentional binding by monetary gains and losses. *PloS One*, *7*(12),
971 e53421.

972 Tsakiris †, M., & Haggard, P. (2005). Experimenting with the acting self. *Cognitive Neuropsychology*,
973 *22*(3–4), 387–407. https://doi.org/10.1080/02643290442000158

974 van der Wel, R. P. R. D. (2015). Me and we: Metacognition and performance evaluation of joint actions.
975 *Cognition*, *140*, 49–59. https://doi.org/10.1016/j.cognition.2015.03.011

976 Vanlangendonck, F., Takashima, A., Willems, R. M., & Hagoort, P. (2018). Distinguishable memory
977 retrieval networks for collaboratively and non-collaboratively learned information. *Neuropsychologia*,
978 *111*, 123–132. https://doi.org/10.1016/j.neuropsychologia.2017.12.008

979 Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for
980 psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*(3), 158–189.

981 Wen, W. (2019). Does delay in feedback diminish sense of agency? A review. *Consciousness and*
982 *Cognition*, *73*, 102759. https://doi.org/10.1016/j.concog.2019.05.007

983 Wen, W., Yamashita, A., & Asama, H. (2016). Divided Attention and Processes Underlying Sense of
984 Agency. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00035

985 Wenke, D., Fleming, S. M., & Haggard, P. (2010). Subliminal priming of actions influences sense of control
986 over effects of action. *Cognition*, *115*(1), 26–38. https://doi.org/10.1016/j.cognition.2009.10.016

987 Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in
988 which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology:*
989 *General*, *143*(5), 2020.

990 Wolpert, D. M., Doya, K., & Kawato, M. (2003). A unifying computational framework for motor control
991 and social interaction. *Philosophical Transactions of the Royal Society of London. Series B, Biological*
992 *Sciences*, *358*(1431), 593–602. https://doi.org/10.1098/rstb.2002.1238

993

994