Susceptibility of agency judgments to social influence

Axel Baptista, Pierre O Jacquet, Nura Sidarus, David Cohen, Valérian Chambon

Susceptibility of agency judgments to social influence

Axel Baptista (1,2,3), Pierre O. Jacquet (1,5,7), Nura Sidarus (5,6), David Cohen (2,4), Valérian Chambon (1).

(1) Institut Jean Nicod, Département d'études cognitives, ENS, CNRS, PSL University, 75005 Paris, France

(2) Service de Psychiatrie de l'Enfant et de l'Adolescent, GH Pitié-Salpêtrière Charles Foix, APHP, Paris, France.

(3) Université de Paris, France.

(4) Institut des Systèmes Intelligents et de Robotique, Sorbonne Université, ISIR CNRS UMR 7222, Paris, France.

(5) Laboratoire de Neurosciences Cognitives & Computationnelles, Département d'études cognitives, École normale supérieure, INSERM, PSL University, 75005 Paris, France

(6) Department of Psychology, Royal Holloway University of London, Surrey, United Kingdom of Great Britain and Northern Ireland.

(7) Institut du Psychotraumatisme de l'Enfant et de l'Adolescent, Centre Hospitalier de Versailles et Conseil départemental des Yvelines et des Hauts de Seine, Versailles, France.

Correspondence can be directed to Axel Baptista, Institut Jean Nicod, Ecole Normale Supérieure, 29 rue d'Ulm 75005 Paris, France (email: axel.baptista@gmail.com; ORCID : https://orcid.org/0000-0002-0780-5755)

ABSTRACT

The experience of agency refers to the phenomenal experience of being the causal source of one's own actions, and through them, the course of events in the outside world. This experience is crucial for the production of adaptive actions, and for the adequate communication of felt action control to peers. The present study examines the possibility that, on certain occasions and under specific internal and external constraints, people rely on explicit social information provided by their peers to revise their self-reports of perceived control, i.e., their judgment of agency. To test this hypothesis, we adapted a task based on an interactive computer game. We manipulated well-known sensorimotor agency cues related to action control, as well as social information communicated to participants by two advisors. We measured the contribution of social and non-social sources of information to agency judgments. We found that at the single-trial level, participants align their JoA with advisor feedback based on their own performance during the task, the type of feedback provided by advisors, and the interaction of this social feedback with the sensorimotor agency cues. At the same time, JoA alignment in previous trial also predicted participants' tendency to revise their JoA after social feedback. Overall, these results demonstrate that agency judgment is subject to social influence. This influence is the result of the integration of social and non-social information at the scale of a single judgment, while also being driven by repeated past interactions with peers.

Keywords: Sense of agency; Social influence; Advice taking; Change-of-mind; Performance; Sensorimotor cues; Serial dependence.

1. Introduction

The experience of agency, also referred to as “Sense of Agency” (SoA), is classically defined as the phenomenal experience of being the causal source of one’s own actions, and through them, the course of events in the outside world (Haggard, 2017; Haggard & Chambon, 2012; Synofzik et al., 2008). A functional sense of agency can serve the useful purpose of having a correct estimate of one’s degree of control in a particular context. For example, it is crucial for an airline pilot to have a correct estimate of their degree of control over the aircraft being flown. However, the mere feeling of control of the aircraft is not enough for the pilot to determine whether they are in control or not. In this situation, the co-pilot’s input can be useful to update the pilot’s possibly inaccurate estimate of perceived control.

Under certain circumstances, information provided by peers can be critical for the production of appropriate actions (Kendal et al., 2009; Toelch et al., 2014). Yet, the conditions under which individuals integrate this ‘social’ information to revise their own sense of agency remains poorly characterized. Indeed, most studies have focused on the contribution to SoA of internal information that encompasses various agency cues related to different stages of action control (e.g., sensory feedback prediction, efferent motor commands, observed sensory feedback, see (Blakemore et al., 2002; Sidarus, Vuorre, & Haggard, 2017; Sidarus, Vuorre, Metcalfe, et al., 2017). In contrast, few studies have examined the integration of these internal sensorimotor cues with information from the social context (Beyer et al., 2017, 2018; Dewey et al., 2014; Sidarus et al., 2020).

The aim of the present study was to examine whether individuals – on certain occasions and under specific internal and external constraints – can rely on explicit social information provided by their peers to revise their own sense of agency (SoA). This suggestion is supported by various lines of research and empirical work. First, modern theoretical frameworks of SoA emphasize that a sense of agency results from the optimal integration of information from multiple sources, including (i) internal information related to the preparation and control of action, and (ii) external information related to the context of the action (e.g., social information) (Gallagher, 2012; Synofzik et al., 2013). Moreover, a significant number of studies has shown that human subjects show some sensitivity to social information, and occasionally modify their decisions or behaviours to match those of others (Jacquet et al., 2018, 2019; Olsen et al., 2019; Pescetelli et al., 2021; Pescetelli & Yeung, 2020). In principle, the use of social information allows individuals to benefit from solutions that have already been tried out by their peers, and is driven by the motivation to achieve an accurate representation of the world (but see Morin et al., 2021 for a recent discussion).

To examine the influence of social and non-social sources of information on the SoA, we designed a computer-based experiment that mixed well-validated paradigms specifically developed to manipulate internal sensorimotor agency cues (Metcalfe et al., 2010, 2012; Metcalfe & Greene, 2007; Sidarus, Vuorre, Metcalfe, et al., 2017) and third-party social information (Jacquet et al., 2019). On each trial in our experiment, we asked participants to play a game in which sensorimotor cues were carefully manipulated to generate different levels of correspondence between intentions, actions, and action outcomes. Specifically, the game involved moving a mouse cursor (a box) along a horizontal bar to catch as many falling X’s as possible, while avoiding distractors (O’s). Participants’ motor actions produced two levels of outcome: *proximal* action outcomes, represented by the movements of the cursor on the screen, and

distal action outcomes, represented by catching X's and making them disappear from the screen (Sidarus, Vuorre, Metcalfe, et al., 2017). Proximal and distal action outcomes were manipulated in two game conditions labelled "Turbulence" and "Magic", respectively. Then, on each trial, participants explicitly reported their SoA corresponding to their sense of control over the cursor during the game.

After this initial Judgment of Agency was produced (hereafter referred to as JoA#1), participants received feedback from virtual advisors. This 'social' feedback consisted of the advisor's inference of the participant's degree of control over the game. We experimentally manipulated the valence and strength of the social feedback so that it could deviate positively or negatively from the participant's initial JoA to a small, medium, and large extent (disagreement trials), or could not deviate at all (agreement trials). In other word, the advisor could believe that the participant's degree of control over the game was less than, greater than, or equal to what they openly reported in the initial JoA. Importantly, the content of the 'social' feedback on each trial was conditional on the participants' initial JoA. After receiving social feedback, participants were asked to report their final, potentially revised JoA (hereafter referred to as JoA#2). Finally, we investigated whether the alignment of the participants' JoA with advisor feedback could be modulated by social preferences. To this end, we manipulated the facial traits of the advisors on the dominance (dominant vs non-dominant facial traits) and trustworthiness dimensions (trustworthy vs untrustworthy facial traits) (Osterhoof & Todorov, 2008).

Our main hypothesis was that participants would revise their JoAs based on advisor feedback **(hypothesis #1)**. Consistent with a number of existing studies, we expected this main effect to be modulated by several experimental factors, such as the strength of advisor feedback **(hypothesis #2)** (Jacquet et al., 2018, 2019), the value of advisor feedback and its interaction with the game condition **(hypothesis #3)**, task performance (Morgan et al., 2012; Toelch et al., 2009) **(hypothesis #4)** and baseline feeling of control **(hypothesis #5),** and finally, ostensive communicative cues provided by virtual advisors' facial traits of dominance and trustworthiness **(hypothesis #6)** (Mercier 2020; Sperber 2010; Todorov 2015; Safra 2017). The rationale for each hypothesis is presented in more detail in Section 3.3.

As an exploratory hypothesis, we tested whether past interactions between participants and advisors – that is, the history of disagreements and agreements between the two parties – impacted the JoAs' alignment with advisors' feedback. We expected to observe a 'reciprocity' effect, whereby receiving an agreement from an advisor on a previous trial increased the likelihood of aligning the JoA with a disagreement from that same advisor on the current trial (Pescetelli et al., 2021). Finally, we also tested whether participants' prior JoA alignment influenced subsequent social information use.

2. Methods

2.1. Participants

Multilevel regression models were used to analyse the data (Gelman & Hill, 2006). However, conventional power calculations are difficult to perform for these models because of the multiple sources of variance that must be taken into account (Westfall et al., 2014). In the absence of existing data with regard to our research goal, sample size was determined *a priori* based on previous studies on social information use (Olsen et al., 2019; Pescetelli & Yeung, 2020), and constrained by participants availability. Given these elements, we aimed for a target sample size of 40 participants. Five participants were discarded due to technical error. All participants (21 females, mean age = 23.61, SD = 3.22) were right-handed, with normal or corrected-to-normal vision, and neurologically healthy. The study was approved by the local Ethical Committee (CER-Sorbonne Université n° 2019–CER2 SOTIPAD), and was carried out in accordance with the Declaration of Helsinki (World Medical Association, 2013). All participants provided informed consent and received payment for participating in the study, and reported being naive to the purpose of the experiment.

2.2. Experiment

Participants carried out 3 experimental blocks of trials (one “non-social” block and two “social” blocks; see *General procedure* below) in the presence of the experimenter. The experimental blocks were implemented using Psychtoolbox 3 (www.psychtoolbox.org) in MATLAB (MathWorks Inc.). On each trial of each block, participants played a computer game whose basic procedure and instructions were adapted from a previous study (Metcalfe & Greene, 2007). Briefly, participants were instructed to move a white box (the cursor) along a grey horizontal track (the *proximal* action outcome) to catch downward falling cross-shaped visual stimuli (X) on the screen and, at the same time, avoid touching disc-shaped stimuli (O) (the *distal* action outcome) (see *Figure 1*).

Manipulation of sensorimotor cues. Three “agentive” game conditions were designed with the aim of manipulating internal sensorimotor agency cues by altering the control that participants exert over the cursor via the mouse. In the CONTROL condition, the objective control of the cursor by the mouse was undistorted (i.e. the player was objectively in full control). In the TURBULENCE condition, objective control was impaired by turbulences (random noise) intervening between the mouse position and the cursor position. Finally, in the MAGIC condition, the radius that would count for a ‘hit’ was extended such that the participant was credited with touching an X even if they had not touched it (Metcalfe et al., 2010). The X’s or O’s disappeared as soon as the participant caught them, but continued to fall below the grey horizontal track when the participant failed to catch them. A “beep” sound signalled hits and a “boop” signalled false alarms (i.e., catching an O instead of an X). After 8 secs of play, participants were asked to indicate their JoA (i.e. the degree of control they felt over the cursor, on a Likert scale from 1, "No control", to 7, "Full control") (Sidarus, Vuorre, Metcalfe, et al., 2017). The scale remained on the screen until participants selected a value. Prior to the task, participants were informed that they would be asked to indicate, on each trial, their degree of control over the cursor on a scale from 1 to 7. Participants were also trained to both play the game and make agency judgments using the Likert scale.

Manipulation of social information. In the two “social” blocks, participants performed the agentive task described above and then received feedback from an "online player" (the advisor) about their degree of control during the game (see *Figure 1*). Advisors were, in reality, bogus agents designed to randomly agree, while adhering to the following constraints: the advisor's rating was consistent with the participants' initial JoA (hereafter referred to as JoA #1) in one third of the trials, and inconsistent in the remaining two thirds. In other words, an advisor could communicate via their rating that the participant was more or less in control than what they expressed via their own rating (disagreement trials), or was consistent with what they ‘perceived’ from watching their action during the game (agreement trials). Disagreement trials were equally split into six possible outcomes: the advisor’s ratings could be either higher (positive disagreement) or lower (negative disagreement) than the participant’s initial ratings; and positive and negative deviations could be either small (+1/-1 point deviation), moderate (+2/-2 points deviation) or large (+3/-3 points deviation). In summary, disagreement trials varied in terms of the *valence* (positive vs. negative) and the *strength* of the disagreement (small vs. moderate vs. large). The valence of the disagreement could be reversed based on participants’ JoA #1, to ensure that the feedback was within scale limits (between 1 and 7). Previous studies using a similar procedure report that participants are confident that group evaluations are provided by real individuals (e.g. Campbell-Meiklejohn et al., 2010; Jacquet et al., 2018, 2019; Klucharev et al., 2009).

Manipulation of first impressions. The advisors were represented by facial avatars generated using FaceGen Modeller 3.5, based on methods developed by Oosterhof and Todorov (Oosterhof & Todorov, 2008). A total of 4 avatars were used for the experiment. Their facial traits varied along (i) the dominance dimension (108 trials, “dominance” pairs in which one advisor – i.e. one facial avatar – was represented by a face with predominantly dominant traits, whereas the other with predominant non-dominant traits), and (ii) the trustworthiness dimension (108 trials, “trustworthiness” pairs in which one advisor was represented by a face with predominantly trustworthy traits, whereas the other was represented with predominant untrustworthy traits), with the aim of manipulating the participants’ first impressions about the advisors’ social intentions. The two advisors had masculine facial traits and represented a Caucasian phenotype. The first reason is that computerized faces are bold (a physical feature which is more likely in males) in order to facilitate the detection of facial expressions. The second reason is that the perception of dominance and untrustworthiness depends on morphological features that are more widely distributed in the male population (Oosterhof & Todorov, 2008). The third reason is that we wanted to avoid variation in stereotypes that could affect advice taking. Trials in which the participants interacted with avatars varying on either the dominant dimension or the trustworthiness dimension were distributed into two distinct “social” blocks. In each of these “social” blocks, participants interacted in a sequential way with two advisors who randomly displayed their feedback (e.g., in trial *x* the feedback was provided by a “dominant” advisor, in trial *y* the feedback was provided by a “non-dominant” advisor). After each advisors’ feedback, participants were asked to give a second control rating (second JoA, hereafter referred to as JoA #2) before moving on to the next trial (see *Figure 1*).

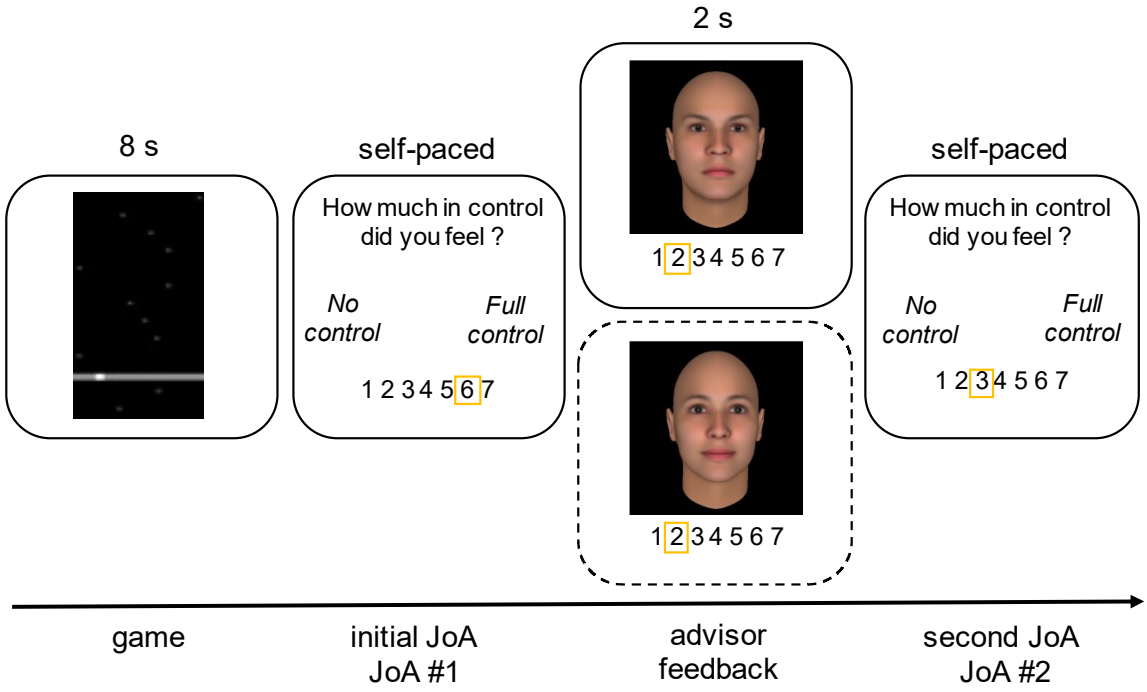


FIGURE 1 | Typical trial of the "social" block: before they were asked to give a second control rating, feedback was given to participants by one of the two advisors. The extent to which participants were influenced by social feedback was estimated by comparing the first (JoA #1) and second (JoA #2) JoA. In this example, the feedback provided by the advisor (top advisor, solid line) substantially changes the participant's JoA, from 6 (JoA #1) to 3 (JoA #3). The dashed line around the bottom avatar in this example meant to indicate that that face was not shown on the trial represented.

General procedure. The experimental sessions were conducted on the experimental platform of the INSEAD-Sorbonne Université Behavioural Lab. The participants were tested in a dedicated room including a computer connected to the local network by a cable visible to the participants. Each participant was brought individually in front of the computer, and was equipped with noise-cancelling headphones.

The experimental protocol was carried out in two sessions on two different days in order to reduce the effect of fatigue, and separated by a period of at least three days. Each session lasted approximately 45 minutes. The first session started by a short training of 18 trials. Then, the experiment started and the participants had to complete a short "non-social" block where each of the 18 trials that composed it consisted in playing the game and producing JoAs. This "non-social" block allowed us to measure the participants' baseline JoAs, i.e. JoAs that could be influenced by the sensorimotor cues (Turbulence and Magic) and task performance only, and not by social information or social preferences. After having completed this "non-social" block, participants were asked to freely choose a nickname and select a facial avatar among 9 possible computerized faces drawn from the Todorov's database. They were then asked to complete consecutively two distinct "social" blocks of 108 trials each (i.e. 216 trials in total): one displaying the "dominance" pairs and the other displaying the "trustworthiness" pairs. The order of the two social blocks was counterbalanced across participants. In each of them, participants performed 36 trials of each "agentive" game condition (Turbulence or Magic or Control). Importantly, after every 9 trials of these

"social" blocks, participants were presented with an 8 seconds duration movie featuring the advisor playing the game and producing a JoA (observational trials). In each movie, the simulated performance and control of the advisors over the cursor were varied in a similar way as it was varied for the participants. Then, participants had to rate advisor's degree of control over the cursor. These trials were not included in the total trial count and were not analyzed. The rationale for presenting these observational trials was to make the online game more realistic and to further convince the participant that the social feedback was provided by real advisors. We conducted a verbal debriefing with the participants after the task. None reported having any doubt that the avatars and their judgements were made by real human agents.

3. Data analyses

Data were analysed using the lme4 (Bates et al., 2015) package in R version 3.4.1 (Team, 2014). All fixed effects of our statistical models were allowed to vary between participants (i.e., participants' uncorrelated random intercepts and slopes). Regression coefficients (β) of the models of interest, their associated statistics, as well as bootstrapped 95% confidence intervals, are reported. We checked for multicollinearity by calculating a variance inflation factor (VIF) for each regressor (Shieh & Fouladi, 2003) (see *Supplementary Material* for details).

Dependent variables. Statistical analyses were performed on three dependent variables, collected on a trial-by-trial basis: *Performance*, *JoA #1*, and *JoA alignment.*

Performance was assessed, on each trial, by calculating a hit rate provided by the number of touched X's divided by the sum of the number of touched X's and missed X's) (see *Supplementary Material* for further analyses on d-prime scores).

As specified above, *JoA #1* corresponds to the initial judgment of agency (from 1, "No control", to 7, "Full control") produced right after the game, whereas *JoA alignment* corresponds to all cases where participants adjusted their JoA #2 (i.e., the second JoA produced after the presentation of the advisors' feedback) in line with the information provided by the advisor. JoA alignment was calculated across all disagreement trials (large, moderate, and small positive/negative disagreements), and was coded as follows: *No JoA alignment* = 0, *JoA alignment* = 1. JoA alignment is considered a proxy of a participant's susceptibility to social influence.

3.1. Model 1 ($\mathcal{M}_1$). Effect of game conditions on performance.

As preliminary analyses, we first checked the ability of our experimental paradigm to correctly manipulate the internal agency cues, in line with previous research (Metcalfe et al., 2010, 2012; Metcalfe & Greene, 2007; Sidarus, Vuorre, Metcalfe, et al., 2017). For this, we used the following linear regression:

$$Performance = \beta_0 + \beta_1.Turb + \beta_2.Magic + \gamma.Z + \varepsilon$$

where $\gamma.Z$ is the random term; *Turb* (deviation coding: turbulence = 1, control = -1, magic = 0) and *Magic* (deviation coding: magic = 1, control = -1, turbulence = 0) represent the agentive game conditions, and *Performance* is the dependent variable.

It is noteworthy that game conditions were "deviation coded" in this model as well as in the following (see $\mathcal{M}_2$ and $\mathcal{M}_3$ below). As such, the parameter estimate (β) for the Turbulence condition is the mean of the dependent variable for Turbulence condition minus the mean of the dependent variable for the Control and Magic conditions combined. Likewise, the β for the Magic condition is the mean of the dependent variable for the Magic condition minus the mean of the dependent variable for the Control and Turbulence conditions combined. The intercept (β_0) is the grand mean. The β of each variable – and its associated t-tests (t, p), calculated using the Satterthwaite approximation for degrees of freedom (Kuznetsova et al., 2017) – represents its independent contribution to trial-by-trial fluctuations in Performance. We expected *Turb* and *Magic* to have a significant negative and positive effect on *Performance*, respectively.

3.2. Model 2 ($\mathcal{M}_2$). Non-social predictors of JoA #1.

We then performed a second preliminary analysis on the JoA#1 ratings. Instead of modelling JoA #1 ratings with a linear model, we treated them as a proportion of the maximum JoA #1 rating (i.e. 7), using the same method as Sidarus and colleagues (Sidarus, Vuorre, & Haggard, 2017). This transformation of the dependent variable was made possible by the fact that (i) JoAs ratings were bounded at 1 and 7, and (ii) a logistic model treating JoA #1 as a proportion predicted JoA #1 better than a linear model (see *Supplementary Material*; *Figure S1* for details), suggesting a curvilinear relationship between JoA #1 and their predictors (see also Figure 3B). For this purpose, we used the following logistic regression:

$$logit\left({}^{JoA\ \#1}/_{7}\right) = \beta_0 + \beta_1.Turb + \beta_2.Magic + \beta_3.Perf$$
$$+ \beta_4.Turb * Perf + \beta_5.Magic * Perf + \gamma.Z + \varepsilon$$

where $\gamma.Z$ is the random term; *Turb* (deviation coded: turbulence = 1, control = -1, magic = 0) and *Magic* (deviation coded: magic = 1, control = -1, turbulence = 0) represent the agentive game conditions; *Perf* refers to the effects of the hit rate (standardized between participants); *Turb*Perf* and *Magic*Perf* represent the interaction terms between the game conditions and performance during the game; and *JoA#1* is the dependent variable. The parameter estimate (β) of each variable, and its associated statistics (z-scores and p-values), calculated using the Wald method (Bates et al., 2015), represents its independent contribution to the trial-by-trial fluctuations in JoA #1.

We expected *Turb* to have a significant negative effect on *JoA #1.* We also expected *Magic* to have a significant negative effect on *JoA #1.* We further explored whether, in the Magic condition, participants' JoA was less influenced by their performance than in the other game conditions. A decrease in the effect of performance on *JoA #1* in this condition would mean that participants are more aware of the artificially enhanced disappearance of the touched X's (i.e. they are more aware of their lack of control) when performance is high than when it is low. If this is the case, then *Magic*Perf* should have a significant positive effect on *JoA #1.*

3.3. Model 3 ($\mathcal{M}_3$). Social and non-social predictors of JoA alignments.

After checking that the game conditions had an impact on participants' performance and judgments of agency (JoA #1) in the expected directions, we tested our main hypotheses, namely: whether and how individuals revise their JoA based on advisors' feedback (**hypothesis #1**), and what sources of information – social or non-social – contribute to JoA revision (**hypothesis #2 to #6**). To this aim, we used the following logistic regression:

$$logit(JoAs\ alignment) = \beta_0 + \beta_1.Turb + \beta_2.Magic + \beta_3.Perf$$
$$+\beta_4.JoA\ \#1 + \beta_5.Disagreement\ valence + \beta_6.Disagreement\ strength$$
$$+ \beta_7.Turb * Disagreement\ valence + \beta_8.Magic * Disagreement\ valence$$
$$+ \beta_9.Offline\ mean\ JoA + \beta_{10}.Previous\ alignment + \gamma.Z + \varepsilon$$

where *JoAs alignment* is the dependent variable; $\gamma.Z$ is the random term; *Turb* (deviation coded: turbulence = 1, control = -1, magic = 0) and *Magic* (deviation coded: magic = 1, control = -1, turbulence = 0) represent the agentive game conditions; *Perf* refers to the hit rate (standardized between participants, see *Supplementary Material* for further analyses on d-prime scores); *JoAs #1* controls for a possible regression-to-the-mean effect (Izuma & Adolphs, 2013); *Disagreement valence (*positive = 0.5 vs. negative = -0.5 disagreement) and *Disagreement strength* (strong = 0.5 vs. moderate = 0 vs. low = -0.5) represent the differences between the advisor's feedback and the participant's JoA #1 in the "social" blocks; *Turb*Disagreement valence* and *Magic*Disagreement valence* represent the interaction terms between the game conditions and the valence of the advisor feedback; *Offline mean JoA* represents each participant's mean agency ratings during the "non-social" block and is used as a proxy of the participant's baseline feeling of control; *Previous alignment* represents the participant's behaviour during the previous disagreement trial, when confronted with the same advisor, and was coded as *previous JoAs alignment* = 0.5, *No previous JoAs alignment* = -0.5. The parameter estimate (β) of each variable – and its associated statistics (z-scores and p-values), calculated using the Wald method (Bates et al., 2015) – represents its independent contribution to trial-by-trial fluctuations in JoA alignment.

Hypothesis #1 states that participants align their JoAs on the advisor's feedback. This hypothesis is supported by numerous studies using similar social influence paradigms (Campbell-Meiklejohn et al., 2010; Jacquet et al., 2018, 2019; Klucharev et al., 2009). However, given that in laboratory settings human subjects tend to downweigh social information in favour of personal information even when it is not adaptive to do so (Morin et al., 2021), we expected JoA alignment to occur less than half the time on average in our group of participants.

Hypothesis #2 states that JoA alignment is modulated by the strength of disagreement between participants' initial JoA and the advisor's feedback. Indeed, previous studies showed that participants' reevaluation of their own judgments after receiving peer feedback scales with the strength of peer disagreement (Campbell-Meiklejohn et al., 2010; Jacquet et al., 2018, 2019; Klucharev et al., 2009). We therefore expected that the greater the disagreement with the advisor, the more participants would align their JoAs (β_5).

Hypothesis #3 states that JoA alignment is modulated by the game conditions (β_1, β_2) and their interaction with the valence of disagreement. This hypothesis is based on research (Metcalfe et al., 2010; Metcalfe & Greene, 2007) that reported that individuals are aware of their lack of control over the outcomes of their actions in the Turbulence and Magic game conditions especially. We can therefore predict that in these two game conditions specifically, participants would not align their second JoA when the value of the advisor's feedback is greater than the value of their initial JoA, i.e. when the valence of disagreement is positively signed (β_7, β_8).

Hypothesis #4 states that JoA alignment is modulated by task performance. This hypothesis is motivated by previous studies related to other cognitive domains that have shown that task difficulty is positively related to social information use (Morgan et al., 2012; Toelch et al., 2009). As a result, we expected participants' performance during the game – a well-used proxy for task difficulty – to have a significant negative effect on JoAs' alignment (β_3)**.**

Hypothesis #5 states that participants' baseline feeling of control – here understood as a correlate of performance that is independent of the degree of susceptibility to social influence – predicts their individual tendency to discount social information. This hypothesis is supported by studies that have shown that individual performance is positively correlated with individual feeling of control (Metcalfe et al., 2012; Metcalfe & Greene, 2007) on the one hand, and negatively correlated with the use of social information on the other hand (Morgan et al., 2012; Toelch et al., 2009). We therefore expected that those of the participants who displayed a high baseline feeling of control would be less likely than other participants to align their JoA during the game with the advisor's feedback (β_9).

Hypothesis #6. In three additional extensions of this model (see $\mathcal{M}_{3c,d,e}$ in the *Supplementary Material*), we tested a 6th and final hypothesis stating that the social intentions that participants could infer from advisors' facial traits (i.e. their first impressions) modulate their propensity to take social feedback into account when revising their JoA. This hypothesis originates from the idea that individuals reason about their informant, often unconsciously or implicitly, and infer the trustworthiness of communicated information on the basis of ostensive cues (Mercier, 2020; Sperber et al., 2010). Previous research has focused on several ostensive cues, such as perceived dominance and trustworthiness of others' facial traits (Safra, 2017; Todorov et al., 2015). In particular, previous studies have consistently found that strong leaders (defined as leader perceived as more dominant and less trustworthy) were less likely to be chosen than more trustworthy and less dominant leaders in the general population (Little et al., 2007, 2012; Re et al., 2013; Safra et al., 2017). Based on this literature, we expected that perceived trustworthiness would elicit more JoA alignment than perceived untrustworthiness, and perceived non-dominance compared to perceived dominance.

Exploratory hypothesis. Finally, we investigated whether past interactions between participants and advisors could have an impact on JoA alignment. During the "social" blocks, participants were paired with two advisors. On each trial, one of the two advisors provided feedback to the participant. We therefore investigated whether participants' JoA alignment produced during the current trial correlated with the JoA alignment produced during their previous interaction with the same advisor (β_{10}). We then explored the impact of past agreement on subsequent JoA alignment in an

independent version of the logistic regression model presented above ($\mathcal{M}_{3b}$; see *Supplementary Material* for details).

4. Results

We dropped 6 participants from the analyses, either because they never aligned their agency ratings to the advisor's feedback during the "social" blocks (n = 4), or because they were outliers on objective performance during the game (n = 2; using the 1.5 times interquartile range criterion (Tukey, 1977)). Our final sample thus consisted of 29 participants aged 19 to 31 years old (mean = 23.34, sd = 3.18). Our key results, presented in the following sections, hold even if including the participants who never aligned their agency ratings to the advisor's feedback (see *Supplementary Material Table S1bis & S2bis & S4bis*). In addition, all effects are substantially equivalent with and without including these 4 participants who never aligned (see *Supplementary Material Figure S2*).

4.1. Model 1 *($\mathcal{M}_1$)*. Effect of game conditions on performance

In this section, we report the results of the linear model of performance ($\mathcal{M}_1$; N_{obs} = 6786; N_{subj} = 29); (see *Supplementary Material*; *Table S1* for details). This model included a by-subject random intercept. The agentive game conditions were treated as fixed effects between subjects because the model did not converge when we treated agentive game conditions as a by-subject random effect.

The results showed that the Turb condition had a significant negative effect on Performance ($\mathcal{M}1$; mean hit rate difference between the Turbulence game condition and the other two game conditions = -0.113, β = -0.075, t(27.993)= - 30.69, p < 2e-16, 95% CI = [-0.080, -0.071]; see *Figure 2*). As expected, participants' objective performance (hit rate) during the game was significantly lower in the Turbulence trials compared to the No-Turbulence (combined Magic and Control conditions) trials. In contrast, the Magic condition had a significant positive effect on Performance ($\mathcal{M}1$; mean hit rate difference between the Magic game condition and the other two game conditions = 0.156, β = 0.104, t(27.100) = 46.37, p < 2e-16, 95% CI = [0.099, 0.109]; see *Figure 2*). . Finally, performance was significantly higher in the Magic trials relative to the other trials (trials from the combined Turbulence and Control conditions).

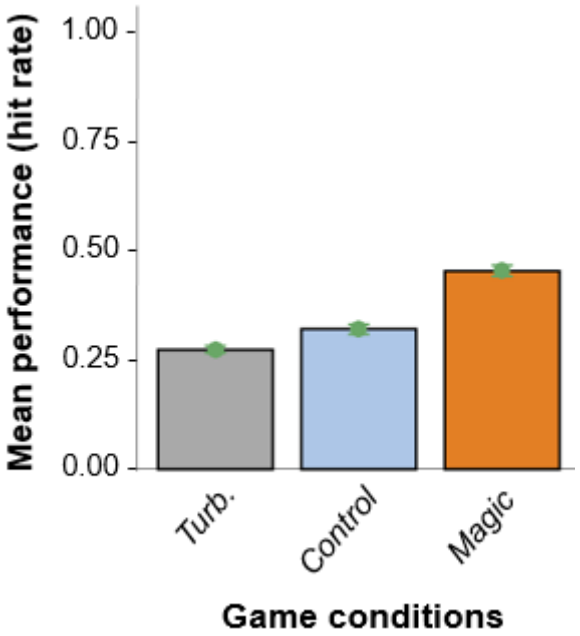


FIGURE 2 | Effect of sensorimotor cues on mean performance ($\mathcal{M}_1$): Effect of sensorimotor cues on mean hit rate. Black error bars show the standard error of the mean. Green error bars show 95% prediction intervals obtained from 10,000 posterior distribution of plausible game conditions parameters values under uniform priors (Gelman et al., 2018).

4.2. Model 2 ($\mathcal{M}_2$). Non-social predictors of JoA #1

In this section, we report the results of the logistic model of JoAs #1 ($\mathcal{M}_2$; N_{obs} = 6786; N_{subj} = 29); (see *Supplementary Material*; *Table S2* for details).

Results revealed that *Perf* had a significant positive effect on *JoA #1* ($\mathcal{M}_2$; β = 0.415, p < 2e-16, 95% CI = [0.347,0.479]; see *Figure 3A & 3B*). As expected, JoAs #1 increased with performance. In addition, results showed that *Turb* had a significant negative effect on *JoA #1* ($\mathcal{M}_2$; β = -0.554, p = 1.63e-13, 95% CI = [-0.703,-0.410]; see *Figure 3A & 3B*), meaning that JoAs #1 were significantly lower in the Turbulence condition. Surprisingly, our results showed that *Magic* had a significant positive effect on *JoA #1* ($\mathcal{M}_2$; β = 0.354, p = 3.43e-14, 95% CI = [0.269,0.450]; see *Figure 3A & 3B*), with participants reporting higher JoAs #1 in the Magic condition. We also found no conclusive evidence that the effect of performance on JoAs #1 was modulated by Magic ($\mathcal{M}_2$; non-significant interaction between Magic and performance: β = 0.011, p = 0.706, 95% CI = [-0.050,0.065]; see *Figure 3A & 3B*). In order to specifically contrast the effect of the Magic condition versus Control condition on JoA #1, we conducted further analysis using the $\mathcal{M}_2$ model including Game agentive conditions (Control, Turbulence and Magic) as a categorical variable after dummy coding (instead of deviation contrast coding), with the Control condition as the reference level. Using this version of the model ($\mathcal{M}_{2b}$), our results showed that JoAs #1 were significantly increased in the Magic versus Control condition (see $\mathcal{M}_{2b}$ in *Supplementary Material; Table S3* for details; β = 0.154, p = 1.21e-4, 95% CI = [0.075, 0.231]). Overall, these results confirm that participants in our study did not report a lack of control in the magic condition, after controlling for performance. On the contrary, participants reported higher JoAs #1 in the Magic condition. We will discuss possible reasons for this finding later (see *Discussion*).

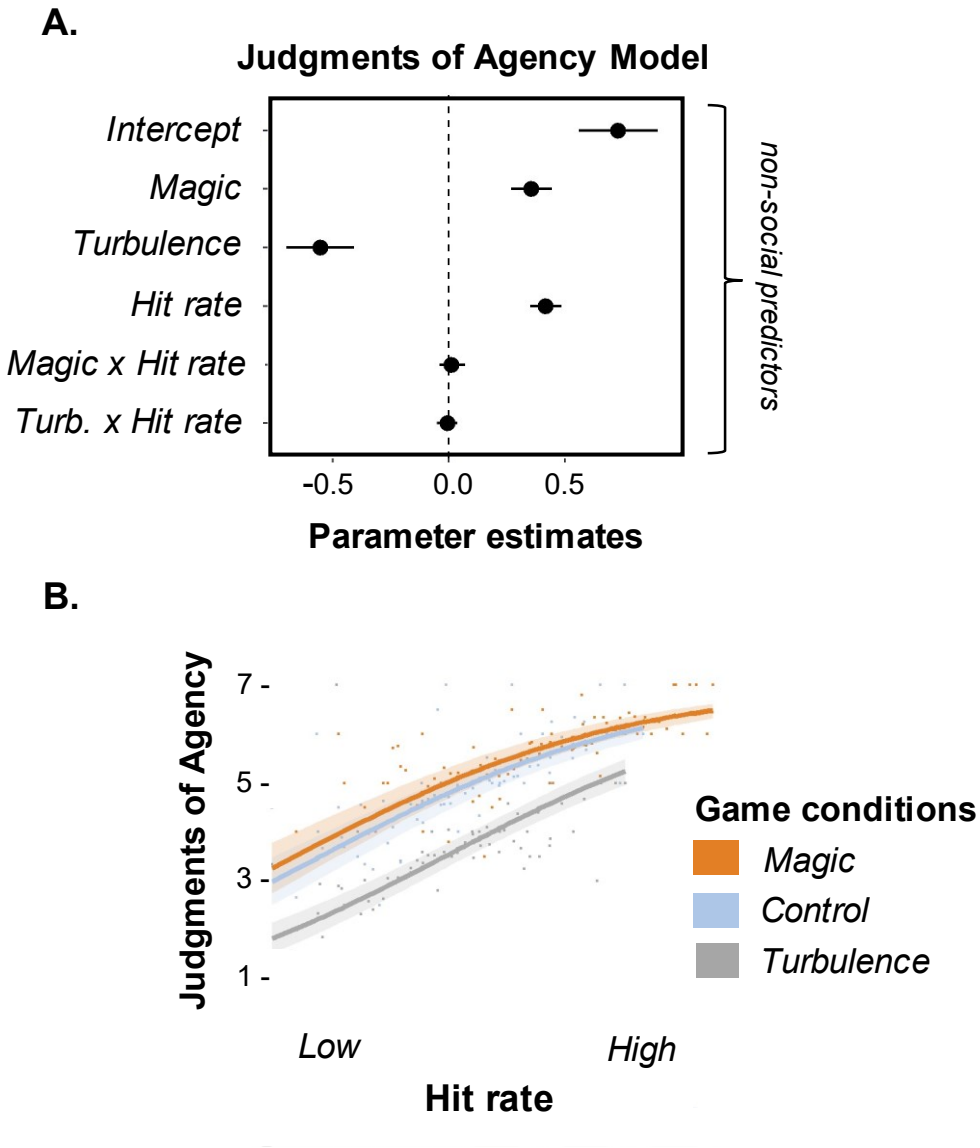


FIGURE 3 | Predictors of JoA #1 ($\mathcal{M}_2$): (A) Parameter estimates, with bootstrapped 95% confidence intervals, from the logistic model of JoA #1 (Performance Z-scored between participants). **(B)** Average JoA #1 across participants (points) and $\mathcal{M}_2$ model predictions (regression lines, and shaded 95% prediction intervals) for the relation between the effect of Game conditions on JoA #1 and participants' performance. Predictions were obtained from 10,000 simulations from the posterior distribution of plausible parameter values under uniform priors (Gelman et al., 2018).

4.3. Model 3 ($\mathcal{M}_3$). Non-social and social predictors of JoA alignment

JoA alignment was assessed during the "social" blocks, in which participants received feedback from advisors on their level of control during the game. In this section, the results of the logistic model of JoA alignment ($\mathcal{M}_3$; N_{obs} = 4172; N_{subj} = 29) were reported (see *Supplementary Material*; *Table S4* for details).

Hypothesis #1. Results of the $\mathcal{M}_3$ model first revealed that during these blocks, participants aligned their JoA #2 in the direction of the information provided by the

advisors in 4.1% to 76.3% (depending on the participant) of the disagreement trials. On average, the proportion of JoA alignment was significantly higher than zero (JoAs alignment = 29.167%, S.E.M = 0.033; two-tailed one sample Wilcoxon signed rank test; V-statistic = 435; p-value = 2.697e-06). These results support **hypothesis #1**, which suggests that participants occasionally revise their JoA based on the advisor's feedback. Note, however, that on average participants aligned their JoA less than half the time, regardless of the type of social feedback (proportion of JoA alignments <50%: two-tailed one sample Wilcoxon signed rank test; V-statistic = 407.5; p-value = 4.166e-05). The fact that JoA alignments are less frequent than non-alignments is consistent with previous research showing that individuals downweigh the advice they receive from a social source in favour of their own initial beliefs (for a recent review, see (Morin et al., 2021)).

Hypothesis #2. Our analysis further revealed that the strength of disagreement had a significant and positive effect on *JoA alignment* ($\mathcal{M}_3$; β = 0.597, p = 1.29e-4, 95% CI = [0.289, 0.929]; see *Figure 4A & 4C*), such that participants were more likely to align their JoA with the advisor's feedback when that feedback expressed strong disagreement with their initial JoA #1. Our results also showed that the valence of disagreement had a significant and positive effect on *JoA alignment* ($\mathcal{M}_3$; β = 0.304, p = 0.042, 95% CI = [0.012,0.629]; see *Figure 4A & 4B*). Overall, participants were more likely to align their JoA with advisors when advisors provided a positive disagreement than a negative disagreement (i.e. the advisor judged the participant to be more in control than the participant reported).

Hypothesis #3. We then found that the *Turb*Disagreement valence* interaction term had a significant and positive effect on *JoA alignment* ($\mathcal{M}_3$; significant Disagreement valence by Turbulence condition interaction: β = -0.762, p = 1.41e-08, 95% CI = [-1.031,-0.512]; see *Figure 4A & 4B*). As expected, the effect of Disagreement valence on JoA alignment was weaker in the Turbulence condition than in the other game conditions (see *Supplementary Material section 10.5* for details). Furthermore, we found no conclusive evidence that the game conditions had an effect on JoA alignment ($\mathcal{M}_3$; Turbulence condition: β = -0.062, p = 0.378, 95% CI = [-0.207, 0.077]; Magic condition: β = 0.138, p = 0.074, 95% CI = [-0.010, 0.303]; see *Figure 4A*).

The *Magic*Disagreement valence* interaction term was also significant: the effect of Disagreement valence on JoA alignment was higher in the Magic condition than in other conditions ($\mathcal{M}_3$; significant Disagreement valence by Magic condition interaction: β = 0.675, p = 9.47e-07, 95% CI = [0.414, 0.946]; see *Figure 4A & 4B;* see *Supplementary Material section 10.5* for details).

Hypothesis #4. Consistent with our prediction from hypothesis #4, we found evidence suggesting that participants' propensity to use social information in the context of our task was modulated by its difficulty, as shown by the negative effect of *Perf* on *JoA alignment* ($\mathcal{M}_3$; Hit rate: β = -0.192, p = 0.001, 95% CI = [-0.306, -0.070]; see *Figure 4A & 4C*). This result therefore confirms that performance reduces the alignment of JoA (hypothesis #4).

Hypothesis #5. Examining the effect of the *Offline JoA* variable does not support hypothesis #5. Indeed, this variable had no significant effect on *JoA alignment* ($\mathcal{M}_3$; β = -0.322, p = 0.118, 95% CI = -0.725, 0.073]; see *Figure 4A*). We did not find conclusive evidence that participants who displayed a higher baseline sense of control were less likely than other participants to align their JoA with the advisor's feedback. Nevertheless, the upper bound of the 95% confidence intervals of the parameter

distribution suggests that participants with a higher *Offline JoA* (a higher baseline sense of control) tended to produce a lower JoA alignment in social blocks.

Hypothesis #6. Finally, results from extensions of the $\mathcal{M}_3$ model did not produce conclusive evidence that social intentions derived from participants' first impressions (formed from advisors' facial features) modulated JoAs' alignment (see $\mathcal{M}_{3c,d,e}$ in *Supplementary Material section 10.7* for details).

Exploratory hypothesis. Interestingly, the analysis revealed that *Previous alignment* had a significant and positive effect on *JoA alignment* ($\mathcal{M}_3$; β = 0.214, p = 0.012, 95% CI = [0.033, 0.391]; see *Figure 4A & 4D*). Thus, above and beyond the contribution of non-social and social information at the level of the current trial, participants were more likely to align their JoA #2 with feedback from a particular advisor when they had done so in previous trials. In contrast, we found no conclusive evidence that past agreement had an effect on JoA alignment ($\mathcal{M}_{3b}$; β = 0.119, p = 0.139, 95% CI = [-0.035, 0.281]; see *Supplementary Material; Table S5* for details).

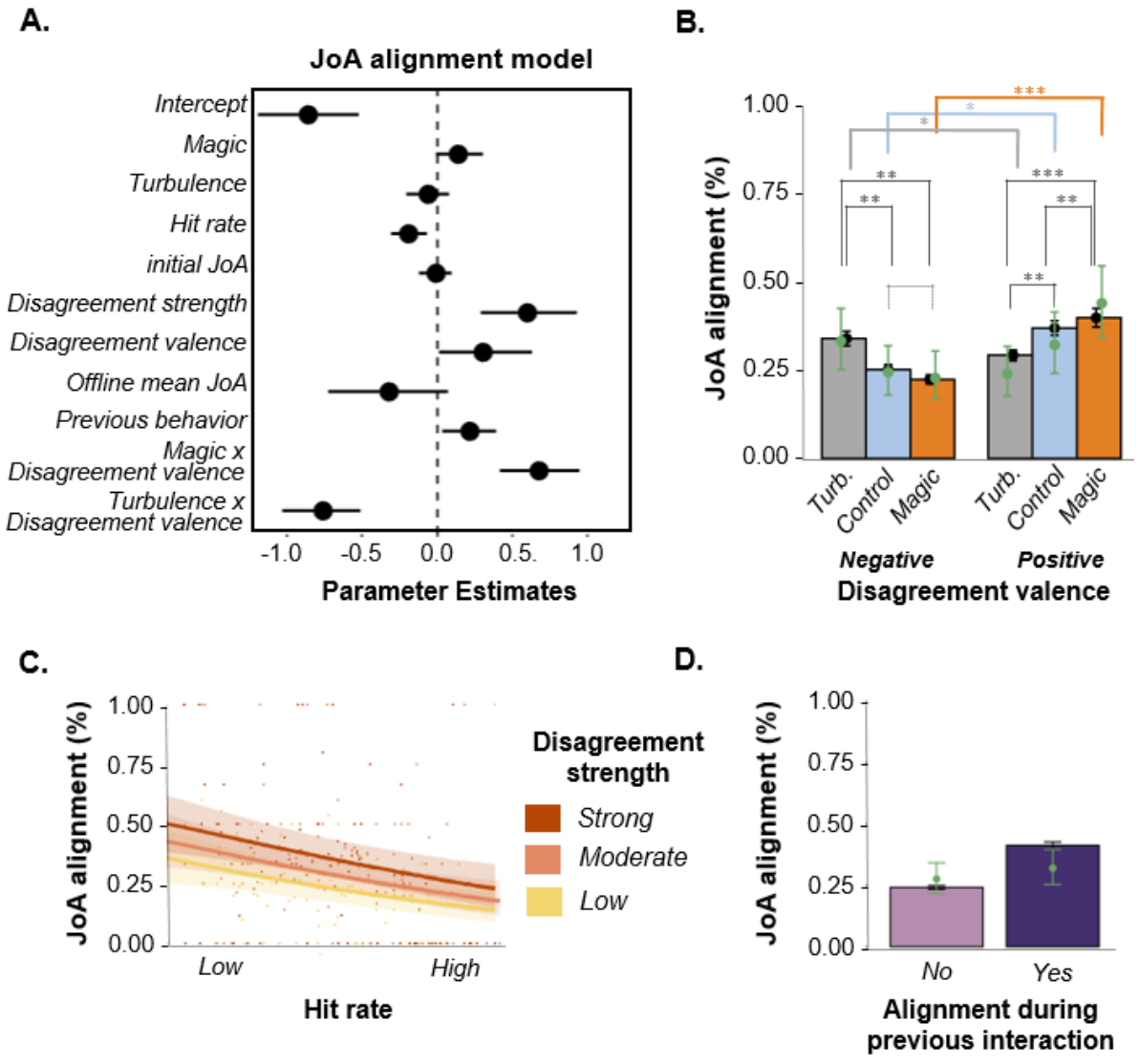


FIGURE 4 | Predictors of JoA alignment ($\mathcal{M}_3$): (A) Parameter estimates, with bootstrapped 95% confidence intervals, from the logistic model of JoA alignment (Hit rate Z-scored between participants). **(B)** Relation between JoA alignment, disagreement valence and game conditions. Black error bars show the standard error of the mean of JoA alignment across participants in the different conditions. Green error bars show 95% prediction intervals obtained from 10,000 simulations from the posterior distribution of plausible *Disagreement valence* and game conditions parameter values under uniform priors (Gelman et al., 2018). Stars indicate significance: * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$ (see *Supplementary Material section 10.5* for details) **(C)** JoA alignment (%) across participants (points) and $\mathcal{M}_3$ model predictions (regression lines, and shaded 95% prediction intervals) for the relation between the effect of Disagreement strength on JoA alignment and participants' Hit rate. Predictions were obtained from 10,000 simulations from the posterior distribution of plausible parameter values. **(D)** Relation between JoA alignment during previous interaction (with the same advisor) and JoA alignment in the current trial (%). Black error bars show the standard error of the mean of JoA alignment across participants. Green error bars show 95% prediction intervals obtained from 10,000 from the posterior distribution of plausible Social context parameter values.

5. Discussion

The aim of the present study was to investigate the conditions under which individuals revise their agency judgments in light of feedback produced by a third party (social information) about their own agentive experience.

In this study, we experimentally manipulated (i) sensorimotor cues related to action control (control vs turbulence vs magic game conditions), (ii) the content of social feedback provided by an external advisor (strength and valence of disagreement), as well as iii) the social preferences evoked by the advisor's facial traits (dominant vs non-dominant; trustworthy vs untrustworthy). We then analysed the effect of these factors on two types of information: participants' performance during the game, and the judgment of agency produced before (JoA #1) and after (JoA #2) the presentation of the advisor's feedback.

Our results reveal that under specific constraints, and given past social interactions, participants rely on explicit social information (the feedback provided by the advisors) to revise their own JoA (see *Figure 5*). This reliance on social information must be contrasted, however, with the fact that in a large majority of trials, participants did not align their JoA with advisors' judgments. This supports the existence, at least in laboratory social learning tasks, of a pervasive phenomenon of egocentric discounting whereby participants downweigh social information in favour of personal information (a phenomenon that has been reported even in task contexts where following social information was more adaptive than ignoring it in favour of individually acquired information, for a recent review, see (Morin et al., 2021)).

5.1 Contribution of non-social information to JoA alignment

We first observe that the alignment of JoAs with advisor feedback is modulated by the non-social information conveyed during the experiment (see *Figure 5*). In particular, participants' performance during the game conditioned the use of social information: participants were less likely to align their JoAs with social feedback when their performance during the game was higher (see model $\mathcal{M}_3$ and Figure 4A & 4C). This result is consistent with previous studies, related to other cognitive domains, that provide converging evidence of increased social information use as a function of task difficulty (Morgan et al., 2012; Toelch et al., 2009). Indeed, JoA depends on objective performance during the task (Metcalfe & Greene, 2007; Sidarus, Vuorre, Metcalfe, et al., 2017). Performance is therefore a reliable cue to consider when evaluating social information about our own JoA. One possible mechanism for this association between performance and social information use is that objective performance during the game (i.e. related to the number of hits accumulated) supports the formation of a context-specific sense of performance. This sense of performance would be exploited to evaluate an information (here, a social information) related to a different cognitive process (the JoA). This possibility supports the debated hypothesis that a sense of confidence based on context-specific information (e.g. performance on a specific task) can be used as a domain-general resource that is recruited to assess information across different cognitive domains (Rouault et al., 2018). Future studies could also examine whether this mechanism also applies to more global estimates of performance over longer time scales.

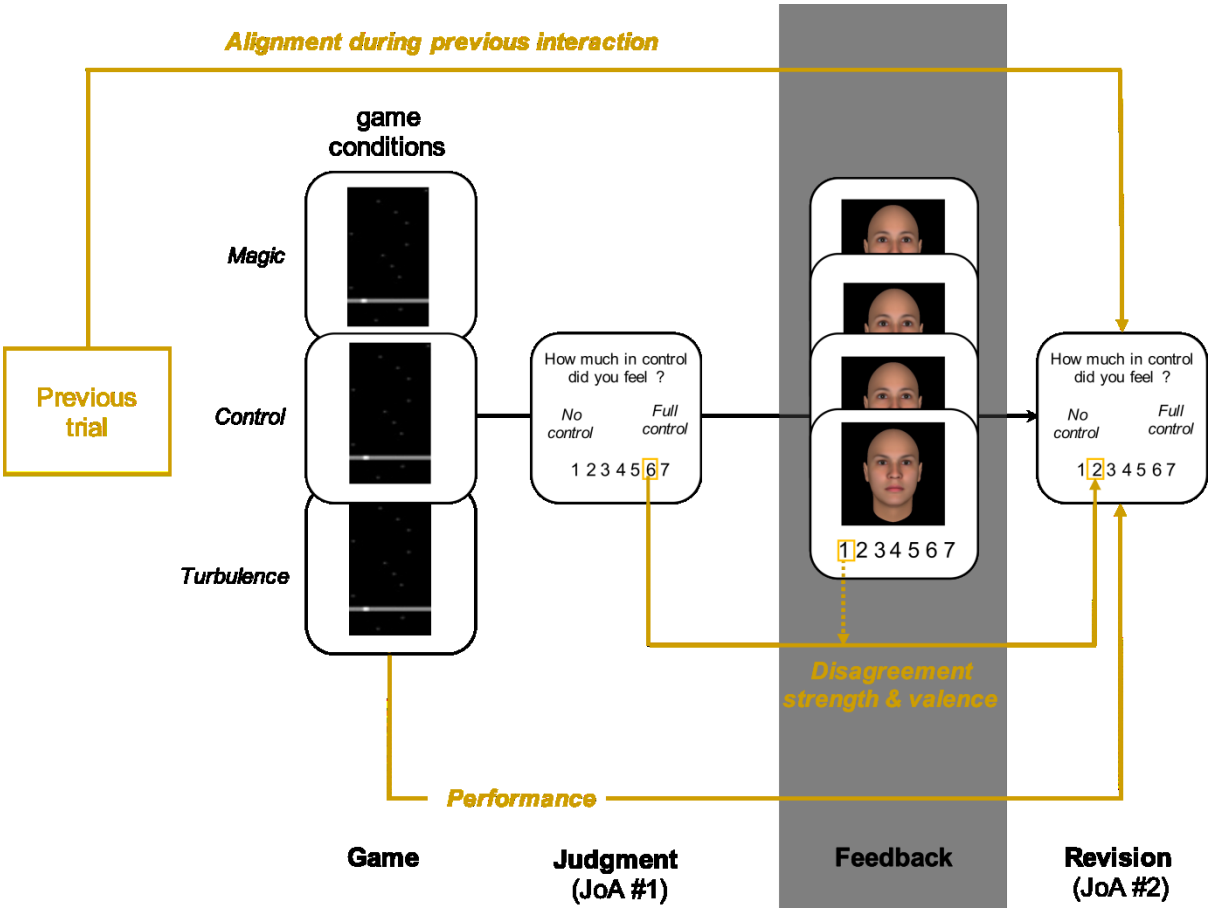


FIGURE 5 | Predictors of JoA alignment. Typical trial of the "social" block with brown arrows representing the different social and non-social predictors of JoA alignment. Interaction effects between predictors are not shown here ($\mathcal{M}3$, see "Results" for details).

5.2 Interplay between non-social and social information

Our study further examined the interaction between non-social information and social information, and the effect of this interaction on participant's behaviour (see *Figure 5*). As mentioned, the task allows for distinguishing between (i) the outcome of a *proximal* action, which corresponds to the movement of a cursor, and which is altered in the Turbulence game condition, and (ii) the outcome of a *distal* action, which corresponds to the disappearance of X's when touched by the cursor, and which is artificially enhanced in the Magic game condition (Sidarus, Vuorre, Metcalfe, et al., 2017). In the latter condition, participants were credited with touching an X even though they did not actually touch it. Our results show that the weight participants place on social information depends on the correspondence between the valence of social information and the disruption of *proximal* action outcomes (Turbulence condition, see Model $\mathcal{M}_3$ and Figure 4A & 4B). Indeed, in the Turbulence game condition, participants discounted social information more when it conflicted with sensorimotor information suggestive of lack of control (i.e. when disagreements were positively signed in the Turbulence game condition) (see model $\mathcal{M}_3$ and Figure 4A & 4B). Conversely, they discounted social information *less* when it agreed with sensorimotor information indicating a lack of control (i.e. when disagreements were negatively signed in the Turbulence game condition) (see model $\mathcal{M}_3$ and Figure 4A & 4B).

Interestingly, in the Control game condition, in which objective mouse control of the cursor was not distorted, participants were more likely to be influenced by advice when disagreements were positive vs. negative (see model $\mathcal{M}_3$ and Figure 4A & 4B). One possible explanation for this result is related to the self-serving bias that predicts a higher sense of agency for desired outcomes (Chambon et al., 2020; Shepperd et al., 2008). We speculate that this “optimistic” bias would further induce a "self-serving advice bias", which makes people more likely to value advice that promotes control over desired outcomes (touching X’s in our study).

We also observed a modulation of the weight given to social information according to the valence of the social information in the Magic condition, which implemented a disruption of the distal action outcome (artificially enhanced in this condition). Indeed, in this game condition, participants were more likely to align their JoA with positive social feedback when it did not match sensorimotor information suggestive of (abnormally) increased control over distal action outcome (i.e. when disagreements were positively signed in the Magic game condition) (see model $\mathcal{M}_3$ and Figure 4A & 4B). In addition, participants relied less on social information when it was consistent with this internal sensorimotor information (i.e. when disagreements were negatively signed in the Magic game condition) (see model $\mathcal{M}_3$ and Figure 4A & 4B). This result may be explained in part by the fact that participants were asked to rate their sense of control over cursor movements (the proximal action outcome), movements that were not disrupted in the Magic condition. Thus, participants had no reason to report a decrease in JoA in this game condition, in which only the distal action outcome was disrupted. It should be noted that in previous studies using the same game, participants were asked about their sense of control “during the game”. As a result, participants accounted for the disruption of the distal action outcome in the Magic condition and, in fact, reported a reduced JoA, after controlling for performance. Future studies are needed to test whether, for the same game condition, participants’ JoA #1, as well as participants’ use of social information, differ depending on the instructions given, and whether their metacognitive processes are involved in evaluating the link between action and proximal action outcome, or between action and distal action outcome.

5.3 Contribution of social feedback to JoA alignment

Interestingly, participants were on average more likely to align their JoA with advisor judgments when those judgments expressed strong disagreements. This result is consistent with previous research using experiments in which a third party provides social feedback (Jacquet et al., 2018, 2019; Klucharev et al., 2009). A classic explanation for these results is that the discrepancy between one’s own judgment and that of one’s advisor induces a conflict that recruits brain areas that play an important role in reward-driven behaviours (Wu et al., 2016). This conflict would therefore motivate individuals to subsequently modify their response (i.e. conform) in order to decrease the discrepancy. However, this account does not explain why, in our study, participants' JoA alignment is conditional on prior non-social agency cues (i.e. performance during the game and game conditions), as discussed above. Similarly, it does not explain the observed egocentric discounting phenomenon already mentioned.

A more plausible explanation, consistent with another line of research, is related to the confirmation bias (Nickerson, 1998). This bias refers to people’s tendency to discount

opinions that contradict their past judgment. A recent study on this bias showed that people are sensitive to the level of confidence with which social information is communicated, provided that this information aligns with their prior beliefs (Kappes et al., 2020). Thus, people rely more on social information when it confirms their prior beliefs, especially when it is expressed with high confidence. This explanation could in theory be sufficient to account for the dependence of participants' JoA alignment on prior information (on whether or not they are in control) and the egocentric discounting phenomenon. It could also explain our result showing that participants' propensity to change their opinion increases with the extent of advice deviation from participants' initial JoA (that is, with the strength of disagreement). If this is the case, participants would treat the strength of advisor disagreement as the level of confidence in the advisors' opinion. This hypothesis remains to be tested.

Among the characteristics of social feedback, we expected that the social preference evoked by the advisor's facial traits (dominant vs non-dominant; trustworthy vs untrustworthy) would also have an impact on social influence. Our results did not provide conclusive evidence of this effect. One possible explanation is that our statistical power may have been insufficient despite (i) using mixed effect models that have greater sensitivity and reliability relative to standard statistical test (e.g. ANOVAs) that do not simultaneously analyse within- and between-subject variability, and (ii) maximizing the number of trials for each facial traits condition (54 trials per facial trait). An a priori power analysis could have informed us about the risk of type II error associated with this effect. As such, the fact that we did not choose the sample size for this study based on a priori power analysis, for the reason mentioned in section 2.1, is a limitation of this study. Another possible explanation is that participants did not pay sufficient attention and/or did not memorize the different avatars they interacted with sequentially. Indeed, participants never interacted with the same advisor more than twice in a row.

5.4 Contribution of past interactions between participants and advisors to JoA alignment

Beyond the contribution of trial-wise (social and non-social) information to JoA, our results reveal that the history of past interactions between participants and advisors also influenced JoA alignment (see *Figure 5*). Previous research has shown that current trial responses are biased by the previous trial response (i.e. serial dependence) in various domains (perceptual decision making, evaluation of stimuli properties, intentional binding, see (Di Costa et al., 2017; Fischer & Whitney, 2014; Liberman et al., 2014). However, it is not known whether the influence of social information on sense of agency (SoA) is also subject to serial dependence. Our study is the first to reveal that prior alignment of participants' SoA (with feedback provided by one of the two advisors) increases the likelihood that they will align their JoA across trials.

Finally, previous research has shown that participant and advisor history of agreement and disagreement impacted the likelihood of aligning the JoA with a disagreement from that same advisor in the current trial (Pescetelli et al., 2021). Our results did not provide conclusive evidence of such a 'reciprocity' effect between participants and each advisor. One possible explanation is that, as mentioned above, participants did not pay sufficient attention to and/or remember the advisors with whom they interacted. Another possibility is that our model is not sensitive enough to detect this

effect. Furthermore, our model only takes into account previous encounters with the same avatar; it did not take into account all past interactions. More refined models of individual behaviour, such as social learning models (Maurer et al., 2018; Olsson et al., 2020), that take into account the entire history of interactions, might allow to study how social preference as well as the history of past interactions influence trust building and the subsequent decision to align with third-party feedback.

6. Conclusions

Our work shows that judgments of agency (JoA) are subject to social influence. Individuals revise their own JoA in the light of (i) feedback produced by a third party (social information) about their own agentive experience, and its interaction with (ii) sensorimotor cues related to action control, and (iii) self-performance during the task. Interestingly, prior JoA alignments increase subsequent JoA revisions. Thus, our main contribution was to demonstrate that the JoA, which is central to sense of self, is constantly shaped by the integration of social and non-social information over time. We also disambiguated the unique contributions of these different sources of information to the formation of a sense of agency in social context. Our results open the possibility of better understanding how the JoA develops and may be involved in self-disorders fuelled by abnormal social interactions, such as in borderline personality disorder.

7. Acknowledgements

A.B. is supported by a fellowship from the *Université de Paris.*

V.C. was supported by the Agence Nationale de la Recherche grants ANR-16-CE37-0012-01 (ANR JCJ) and ANR-19-CE37-0014-01 (ANR PRC). V.C. and P.O.J. were supported by a department-wide grant from the Agence Nationale de la Recherche (Frontiers in Cognition, ANR-17-EURE-0017) and by ANR-10-IDEX-0001-02 PSL (program “Investissements d’Avenir”)

N.S. was supported by a Fyssen Foundation Postdoctoral fellowship.

We are grateful to the *INSEAD-Sorbonne Université Behavioural Lab* for helping us recruit participants.

This is a preprint of an article published in *Cognition*. The final authenticated version is available online at: https://doi.org/10.1016/j.cognition.2022.105173

8. Author contribution

Conceptualization: A.B, V.C, P.O.J; Methodology: A.B, V.C., P.O.J, N.S; Software: A.B; Investigation: A.B; Formal analysis: A.B; Resources: A.B, D.C, V.C; Data Curation: A.B; Writing – original draft: A.B; Writing – review and editing: A.B, V.C, P.O.J, N.S; Visualization: A.B; Supervision: D.C, V.C, P.O.J; Project administration: A.B, V.C; Funding acquisition: A.B, D.C, V.C.

9. References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

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*(1), 138–145. https://doi.org/10.1093/scan/nsw160

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

Campbell-Meiklejohn, D. K., Bach, D. R., Roepstorff, A., Dolan, R. J., & Frith, C. D. (2010). How the Opinion of Others Affects Our Valuation of Objects. *Current Biology*, *20*(13), 1165–1170. https://doi.org/10.1016/j.cub.2010.04.055

Chambon, V., Théro, H., Vidal, M., Vandendriessche, H., Haggard, P., & Palminteri, S. (2020). Information about action outcomes differentially affects learning from self-determined versus imposed choices. *Nature Human Behaviour*, *4*(10), 1067–1079.

Dewey, J. A., Pacherie, E., & Knoblich, G. (2014). The phenomenology of controlling a moving object with another person. *Cognition*, *132*(3), 383–397. https://doi.org/10.1016/j.cognition.2014.05.002

Di Costa, S., Théro, H., Chambon, V., & Haggard, P. (2017). Try and try again: Post-error boost of an implicit measure of agency. *The Quarterly Journal of Experimental Psychology*, 1–28. https://doi.org/10.1080/17470218.2017.1350871

Fischer, J., & Whitney, D. (2014). Serial dependence in visual perception. *Nature Neuroscience*, *17*(5), 738–743. https://doi.org/10.1038/nn.3689

Gallagher, S. (2012). Multiple aspects in the sense of agency. *New Ideas in Psychology*, *30*(1), 15–31. https://doi.org/10.1016/j.newideapsych.2010.03.003

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gelman, A., Su, Y. S., Masanao, Y., Zheng, T., & Dorie, V. (2018). *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models, version 1.10-1*.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1). Wiley New York.

Haggard, P. (2017). Sense of agency in the human brain. *Nature Reviews Neuroscience*, *18*(4), 196–207. https://doi.org/10.1038/nrn.2017.14

Haggard, P., & Chambon, V. (2012). Sense of agency. *Current Biology*, *22*(10), R390–R392. https://doi.org/10.1016/j.cub.2012.02.040

Izuma, K., & Adolphs, R. (2013). Social Manipulation of Preference in the Human Brain. *Neuron*, *78*(3), 563–573. https://doi.org/10.1016/j.neuron.2013.03.023

Jacquet, P. O., Safra, L., Wyart, V., Baumard, N., & Chevallier, C. (2019). The ecological roots of human susceptibility to social influence: A pre-registered study investigating the impact of early-life adversity. *Royal Society Open Science*, *6*(1), 180454. https://doi.org/10.1098/rsos.180454

Jacquet, P. O., Wyart, V., Desantis, A., Hsu, Y.-F., Granjon, L., Sergent, C., & Waszak, F. (2018). Human susceptibility to social influence and its neural correlates are related to perceived vulnerability to extrinsic morbidity risks. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-31619-8

Kappes, A., Harvey, A. H., Lohrenz, T., Montague, P. R., & Sharot, T. (2020). Confirmation bias in the utilization of others' opinion strength. *Nature Neuroscience*, *23*(1), 130–137. https://doi.org/10.1038/s41593-019-0549-2

Kendal, R. L., Coolen, I., & Laland, K. N. (2009). 13. Adaptive Trade-offs in the Use of Social and Personal Information. In *Cognitive ecology II* (pp. 249–271). University of Chicago Press.

Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A., & Fernández, G. (2009). Reinforcement Learning Signal Predicts Social Conformity. *Neuron*, *61*(1), 140–151. https://doi.org/10.1016/j.neuron.2008.11.027

Kock, N., & Lynn, G. (2012). Lateral Collinearity and Misleading Results in Variance-Based SEM: An Illustration and Recommendations. *Journal of the Association for Information Systems*, *13*(7), 546–580. https://doi.org/10.17705/1jais.00302

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13).

Liberman, A., Fischer, J., & Whitney, D. (2014). Serial Dependence in the Perception of Faces. *Current Biology*, *24*(21), 2569–2574. https://doi.org/10.1016/j.cub.2014.09.025

Little, A. C., Burriss, R. P., Jones, B. C., & Roberts, S. C. (2007). Facial appearance affects voting decisions. *Evolution and Human Behavior*, *28*(1), 18–27.

Little, A. C., Roberts, S. C., Jones, B. C., & DeBruine, L. M. (2012). The perception of attractiveness and trustworthiness in male faces affects hypothetical voting

decisions differently in wartime and peacetime scenarios. *Quarterly Journal of Experimental Psychology*, *65*(10), 2018–2032.

Makowski, D., Ben-Shachar, M. S., Chen, S. H. A., & Lüdecke, D. (2019). Indices of Effect Existence and Significance in the Bayesian Framework. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02767

Maurer, C., Chambon, V., Bourgeois-Gironde, S., Leboyer, M., & Zalla, T. (2018). The influence of prior reputation and reciprocity on dynamic trust-building in adults with and without autism spectrum disorder. *Cognition*, *172*, 1–10. https://doi.org/10.1016/j.cognition.2017.11.007

Mercier, H. (2020). *Not born yesterday: The science of who we trust and what we believe*. Princeton University Press.

Metcalfe, J., Eich, T. S., & Castel, A. D. (2010). Metacognition of agency across the lifespan. *Cognition*, *116*(2), 267–282. https://doi.org/10.1016/j.cognition.2010.05.009

Metcalfe, J., & Greene, M. J. (2007). Metacognition of agency. *Journal of Experimental Psychology: General*, *136*(2), 184–199. https://doi.org/10.1037/0096-3445.136.2.184

Metcalfe, J., Van Snellenberg, J. X., DeRosse, P., Balsam, P., & Malhotra, A. K. (2012). Judgements of agency in schizophrenia: An impairment in autonoetic metacognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1391–1400. https://doi.org/10.1098/rstb.2012.0006

Mill, R. D., & O'Connor, A. R. (2014). Question format shifts bias away from the emphasised response in tests of recognition memory. *Consciousness and Cognition*, *30*, 91–104.

Morgan, T. J., Rendell, L. E., Ehn, M., Hoppitt, W., & Laland, K. N. (2012). The evolutionary basis of human social learning. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1729), 653–662.

Morin, O., Jacquet, P. O., Vaesen, K., & Acerbi, A. (2021). Social information use and social information waste. *Philosophical Transactions of the Royal Society B*, *376*(1828), 20200052.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

Olsen, K., Roepstorff, A., & Bang, D. (2019). *Knowing whom to learn from: Individual differences in metacognition and weighting of social information*.

Olsson, A., Knapska, E., & Lindström, B. (2020). The neural and computational systems of social learning. *Nature Reviews Neuroscience*, 1–16. https://doi.org/10.1038/s41583-020-0276-4

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087–11092.

Pescetelli, N., Hauperich, A.-K., & Yeung, N. (2021). Confidence, advice seeking and changes of mind in decision making. *Cognition*, *215*, 104810. https://doi.org/10.1016/j.cognition.2021.104810

Pescetelli, N., & Yeung, N. (2020). The role of decision confidence in advice-taking and trust formation. *Journal of Experimental Psychology: General*.

Re, D. E., DeBruine, L. M., Jones, B. C., & Perrett, D. I. (2013). Facial cues to perceived height influence leadership choices in simulated war and peace contexts. *Evolutionary Psychology*, *11*(1), 147470491301100100.

Rouault, M., McWilliams, A., Allen, M. G., & Fleming, S. M. (2018). Human Metacognition Across Domains: Insights from Individual Differences and

Neuroimaging. *Personality Neuroscience*, *1*, e17. https://doi.org/10.1017/pen.2018.16

Safra, L. (2017). *Using facial cues to produce social decisions. A cognitive and evolutionary approach* [PhD Thesis].

Safra, L., Algan, Y., Tecu, T., Grèzes, J., Baumard, N., & Chevallier, C. (2017). Childhood harshness predicts long-lasting leader preferences. *Evolution and Human Behavior*. https://doi.org/10.1016/j.evolhumbehav.2017.05.001

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

Shieh, Y.-Y., & Fouladi, R. T. (2003). The Effect of Multicollinearity on Multilevel Modeling Parameter Estimates and Standard Errors. *Educational and Psychological Measurement*, *63*(6), 951–985. https://doi.org/10.1177/0013164403258402

Sidarus, N., Travers, E., Haggard, P., & Beyer, F. (2020). How social contexts affect cognition: Mentalizing interferes with sense of agency during voluntary action. *Journal of Experimental Social Psychology*, *89*, 103994. https://doi.org/10.1016/j.jesp.2020.103994

Sidarus, N., Vuorre, M., & Haggard, P. (2017). Integrating prospective and retrospective cues to the sense of agency: A multi-study investigation†. *Neuroscience of Consciousness*, *3*(1). https://doi.org/10.1093/nc/nix012

Sidarus, N., Vuorre, M., Metcalfe, J., & Haggard, P. (2017). Investigating the Prospective Sense of Agency: Effects of Processing Fluency, Stimulus Ambiguity, and Response Conflict. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00545

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*(1), 34.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*(4), 359–393.

Synofzik, M., Vosgerau, G., & Newen, A. (2008). Beyond the comparator model: A multifactorial two-step account of agency. *Consciousness and Cognition*, *17*(1), 219–239. https://doi.org/10.1016/j.concog.2007.03.010

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

Team, R. C. (2014). *R: A language and environment for statistical computing*.

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social Attributions from Faces: Determinants, Consequences, Accuracy, and Functional Significance. *Annual Review of Psychology*, *66*(1), 519–545. https://doi.org/10.1146/annurev-psych-113011-143831

Toelch, U., Bruce, M. J., Newson, L., Richerson, P. J., & Reader, S. M. (2014). Individual consistency and flexibility in human social information use. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1776), 20132864.

Toelch, U., van Delft, M. J., Bruce, M. J., Donders, R., Meeus, M. T., & Reader, S. M. (2009). Decreased environmental variability induces a bias for social information use in humans. *Evolution and Human Behavior*, *30*(1), 32–40.

Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2). Reading, Mass.

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

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, *6*(3), 291–298.

World Medical Association. (2013). World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. *Jama*, *310*(20), 2191–2194.

Wu, H., Luo, Y., & Feng, C. (2016). Neural signatures of social conformity: A coordinate-based activation likelihood estimation meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, *71*, 101–111. https://doi.org/10.1016/j.neubiorev.2016.08.038

10. Supplementary Material

10.1 Comparison of models predicting Performance.

The baseline model is a linear mixed model, which includes game conditions as independent variables. The alternative model is a linear mixed model which corresponds to the baseline model including d-prime score in each participant instead of hit rate. The d-prime measured discrimination performance between targets (X) and distractors (O) (Green & Swets, 1966), and its calculation was adjusted for instances of zero false alarms (Mill & O'Connor, 2014; Snodgrass & Corwin, 1988).

***Model comparison results*.** Bayesian model comparison was performed using frequentist models via the BIC approximation (Makowski et al., 2019). A Bayes factor of less than 1/3 indicates "substantial" evidence in favour of the reference model. A Bayes factor greater than 3 indicates "substantial" evidence in favour of the alternative model (Wetzels et al., 2011). A Bayes factor greater than 150 indicates very strong evidence in favour of the alternative model. The Bayes factor indicates that the two model have the same predictive power. ($BF_{\text{Baseline model vs. Alt. Model}} = 1$).

Table S1 | linear mixed model of Performance (hit rate) ($\mathcal{M}_1$). Regression coefficients of main and interaction effects.

	Estimate (SE)	***t*-value (df)**	***p*-value**	**2.5%**	**97.5%**
Main effects					
Intercept	0.347 (0.005)	67.62 (27.994)	<2e-16	0.337	0.358
Turbulence	-0.075 (0.002)	-30.69 (27.993)	<2e-16	-0.080	-0.071
Magic	0.104 (0.002)	46.37 (27.9999)	<2e-16	0.100	0.109

Note: Columns show parameter estimates, standard errors, t-values and p-values based on the Satterthwaite approximation (Kuznetsova et al., 2017), lower and upper bounds of the bootstrapped 95% Confidence Intervals (bci).

Table S1bis | linear mixed model of Performance (hit rate), including the 4 participants who didn't aligned their agency ratings to the advisor's feedback during the "social" blocks (Nobs = 7722; Nsubj = 33). Regression coefficients of main and interaction effects.

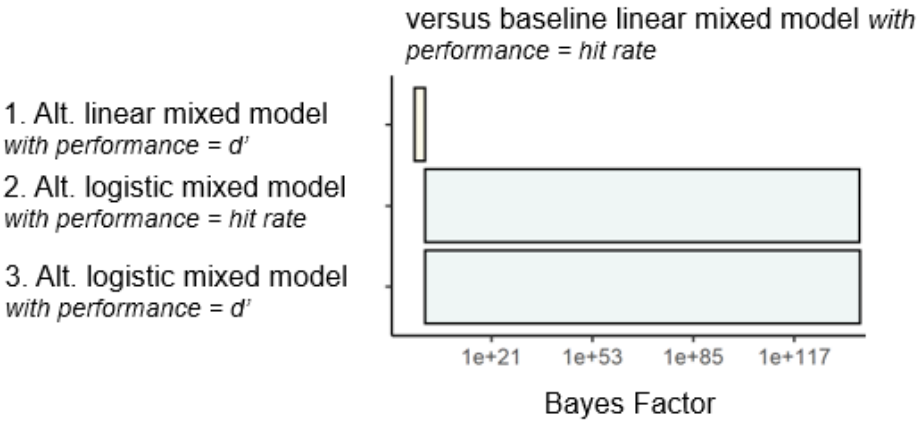
	Estimate (SE)	***t*-value (df)**	***p*-value**	**2.5%**	**97.5%**
Main effects					
Intercept	0.347 (0.006)	58.02 (32.0002)	<2e-16	0.334	0.359
Turbulence	-0.074 (0.002)	-30.93 (32.005)	<2e-16	-0.079	-0.069
Magic	0.104 (0.002)	47.01 (31.997)	<2e-16	0.099	0.108

Note: Columns show parameter estimates, standard errors, t-values and p-values based on the Satterthwaite approximation (Kuznetsova et al., 2017), lower and upper bounds of the bootstrapped 95% Confidence Intervals (bci).

10.2 Comparison of models predicting JoAs #1

The baseline model is a linear mixed model that includes Turbulence, Magic, Hit rate, as independent variables, as well as the interaction between Turbulence and Hit rate variables, as well as Magic and Hit rate. The first alternative model is a linear mixed model that corresponds to the baseline model including d-prime scores calculated for each participant instead of hit rate. The d-prime score measures discrimination performance between targets (X) and distractors (O) (Green & Swets, 1966), and its calculation is adjusted for instances of zero false alarms (Mill & O'Connor, 2014; Snodgrass & Corwin, 1988). The second alternative model is a logistic mixed model including the same independent variables as in the baseline model, and treating JoA as a proportion of the maximum agency ratings (i.e. 7). The third alternative model is a logistic mixed model corresponding to the second alternative model but that includes d-prime scores instead of hit rate. Bayesian model comparison was made using frequentist models via BIC approximation (Makowski et al., 2019). A Bayes factor less than 1/3 indicates "substantial" evidence in favour of the reference model. A Bayes factor greater than 3 indicates "substantial" evidence in favour of the alternative model (Wetzels et al., 2011). A Bayes factor higher than 150 indicates very strong evidence in favour of the alternative model.

Figure S1 | Comparison of models predicting the JoAs #1



Results of model comparison. A general overview (see ***Figure S1***) of the Bayes factors indicates that alternative logistic models treating JoA as a proportion of the maximum agency ratings predicted the observed data better than other linear models. Second-order Bayes factors further indicated non substantial evidence in favour of the alternative logistic model including d-prime scores instead of hit rate $BF_{\text{logistic, d-prime vs. logistic, hit rate}} = 1.522$

Table S2 | Logistic mixed model of JoAs #1 including hit rate ($\mathcal{M}_2$: Nobs = 6786; Nsubj = 29). Regression coefficients of main and interaction effects.

	Estimate (SE)	***z*-value**	***p*-value**	**2.5%**	**97.5%**
Main effects					
Intercept	0.728 (0.081)	8.984	< 2e-16	0.576	0.890
Turbulence	- 0.554 (0.075)	-7.376	1.63e-13	-0.703	-0.410
Magic	0.354 (0.047)	7.581	3.43e-14	0.269	0.450
Performance (Hit rate)	0.415 (0.034)	12.303	< 2e-16	0.347	0.479
Interaction effects					
Turbulence × Performance (Hit rate)	- 0.005 (0.023)	-0.230	0.818	-0.049	0.042
Magic x Performance (Hit rate)	0.011 (0.029)	0.377	0.706	-0.050	0.065

Note: Columns show parameter estimates, standard errors, z-values, p-values, lower and upper bounds of the bootstrapped 95% Confidence Intervals (bci). Parameter estimates standard error and their associated p-values were obtained using Wald method (Bates et al., 2015). "x" denotes interaction terms.

We computed VIF for all the regressors. We found an average VIF of 1.600 s.d. ± 0.227 (median: 1.552, max: 1.891), suggesting that multicollinearity between regressors was not an issue in our model (Kock & Lynn, 2012).

Table S2bis | Logistic mixed model of JoAs #1 including hit rate, and including the 4 participants who (almost) never aligned their agency ratings to the advisor's feedback during the "social" blocks (Nobs =7722; Nsubj = 33). Regression coefficients of main and interaction effects.

	Estimate (SE)	***z*-value**	***p*-value**	**2.5%**	**97.5%**
Main effects					
Intercept	0.697 (0.087)	8.709	< 2e-16	0.533	0.85
Turbulence	-0.569 (0.069)	-8.257	<2e-16	-0.707	-0.43
Magic	0.368 (0.044)	8.415	< 2e-16	0.276	0.454
Performance (Hit rate)	0.406 (0.032)	12.608	< 2e-16	0.345	0.469
Interaction effects					
Turbulence × Performance (Hit rate)	- 0.0003 (0.022)	-0.016	0.987	-0.043	0.0464
Magic x Performance (Hit rate)	0.008 (0.025)	0.312	0.755	-0.041	0.057

Note: Columns show parameter estimates, standard errors, z-values, p-values, lower and upper bounds of the bootstrapped 95% Confidence Intervals (bci). Parameter estimates standard error and their associated p-values were obtained using Wald method (Bates et al., 2015). "x" denotes interaction terms.

10.3 Alternative model of JoAs #1 including the dummy coded "Game conditions" variable ($\mathcal{M}_{2b}$)

Table S3 | Logistic mixed model of JoAs #1 including the dummy coded "Game conditions" variable ($\mathcal{M}_{2b}$;Nobs = 6786; Nsubj = 29). In order to specifically contrast the effect of the Magic versus Control condition on JoA #1, we performed an additional analysis using the $\mathcal{M}_2$ model including the "Game conditions" variable (Control, Turbulence and Magic) as a categorical variable after dummy coding (instead of deviation contrast coding), with the Control condition as the reference level.

	Estimate (SE)	***z*-value**	***p*-value**	**2.5%**	**97.5%**
Main effects					
Intercept	0.928 (0.109)	8.538	< 2e-16	0.702	1.124
Turbulence	- 0.754 (0.110)	-6.888	5.66e-12	-0.962	-0.543
Magic	0.154 (0.040)	3.843	1.21e-4	0.075	0.231
Performance (Hit rate)	0.410 (0.040)	10.175	< 2e-16	0.334	0.490
Interaction effects					
Turbulence × Performance (Hit rate)	0.000 (0.035)	0.003	0.997	-0.071	0.100
Magic x Performance (Hit rate)	0.016 (0.046)	0.359	0.720	-0.070	0.064

Note: Columns show parameter estimates, standard errors, z-values, p-values, lower and upper bounds of the bootstrapped 95% Confidence Intervals (bci). Parameter estimates standard error and their associated p-values were obtained using Wald method (Bates et al., 2015). "x" denotes interaction terms. Game agentive conditions (Control, Turbulence and Magic) were included in the models as a categorical variable after dummy coding with the Control condition as the reference level.

We computed VIF for all the regressors. We found an average VIF of 3.175 s.d. ± 1.294 (median: 2.891, max: 5.103), suggesting that multicollinearity between regressors was not an issue in our model (Kock & Lynn, 2012).

10.4 Comparison of models predicting JoAs alignment

The baseline model is a logistic mixed model that includes Turbulence, Magic, Hit rate, JoA #1, Disagreement strength, Disagreement valence, as main predictors. The model also includes interaction terms between Disagreement valence and game conditions, as well as participant's mean offline JoA and JoAs alignment during the previous interaction with the same advisor. The alternative model corresponds to the baseline model including d-prime score in each participant instead of hit rate. The d-prime score measures discrimination performance between targets (X) and distractors (O) (Green & Swets, 1966), and its calculation is adjusted for instances of zero false alarms (Mill & O'Connor, 2014; Snodgrass & Corwin, 1988). Bayesian model comparison was made using frequentist models via BIC approximation (Makowski et al., 2019). A Bayes factor less than 1/3 indicates substantial evidence in favour of the reference model (Wetzels et al., 2011).

***Model comparison results*.** The models comparison for JoAs alignment indicated that the baseline logistic mixed model treating performance during the game as the Hit rate predicted the observed data better than the alternative model ($BF_{\text{Baseline vs. Alt. Model}} = 0.008 < 1/3$).

Table S4 | Logistic mixed model of JoAs alignment ($\mathcal{M}_3$; Nobs = 4172; Nsubj = 29). Regression coefficients of main and interaction effects of the model are reported in the main text.

	Estimate (SE)	***z*-value**	***p*-value**	**2.5%**	**97.5%**
Main effects					
Intercept	-0.856 (0.162)	-5.263	1.42e-07	- 1.193	- 0.520
Turbulence	-0.062 (0.070)	-0.881	0.38	- 0.207	0.077
Magic	0.138 (0.078)	1.784	0.074	- 0.010	0.303
Performance (Hit rate)	-0.192 (0.056)	-3.413	0.001	- 0.306	- 0.070
JoAs #1	-0.007 (0.056)	-0.134	0.894	- 0.121	0.096
Disagreement strength	0.597 (0.156)	3.829	1.29-e4	0.289	0.929
Disagreement valence	0.304 (0.150)	2.029	0.042	0.012	0.629

Offline mean JoA	-0.322 (0.206)	-1.563	0.118	- 0.725	0.073
Previous alignment	0.214 (0.085)	2.510	0.012	0.033	0.391
Interaction effects					
Magic x Disagreement valence	0.675 (0.138)	4.902	9.47e-07	0.414	0.946
Turbulence x Disagreement valence	-0.762 (0.134)	-5.672	1.41e-08	-1.031	-0.512

Note: Columns show parameter estimates, standard errors, z-values, p-values, lower and upper bounds of the bootstrapped 95% Confidence Intervals (bci). Parameter estimates standard error and their associated p-values were obtained using Wald method (Bates et al., 2015). "x" denotes interaction terms.

We computed VIF for all the regressors. We found an average VIF of 1.519 s.d. ± 0.434 (median: 1.332, max: 2.348), suggesting that multicollinearity between regressors was not an issue in our model (Kock & Lynn, 2012).

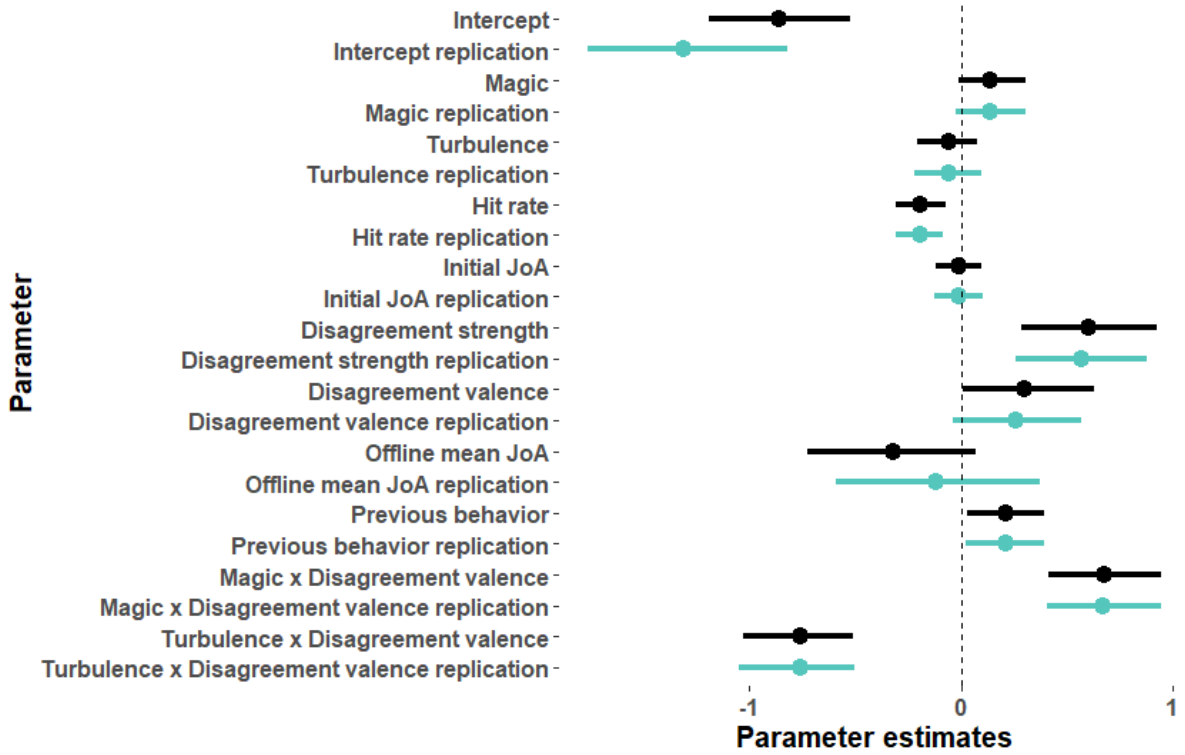
Table S4 bis | Logistic mixed model of JoAs alignment ($\mathcal{M}_3$; Nobs = 4748; Nsubj = 33). Regression coefficients of main and interaction effects of the model are reported in the main text.

	Estimate (SE)	***z*-value**	***p*-value**	**2.5%**	**97.5%**
Main effects					
Intercept	-1.307 (0.243)	-5.389	7.10e-08	-1.762	-0.821
Turbulence	-0.056 (0.077)	-0.728	0.466415	-0.22	0.097
Magic	0.138 (0.083)	1.663	0.096223	-0.024	0.305
Performance (Hit rate)	-0.193 (0.057)	-3.423	0.000619	-0.305	-0.082
JoAs #1	-0.012 (0.056)	-0.207	0.835649	-0.128	0.102

Disagreement strength	0.568 (0.156)	3.632	0.0003	0.261	0.878
Disagreement valence	0.264 (0.146)	1.811	0.07	-0.038	0.572
Offline mean JoA	-0.121 (0.25)	-0.484	0.628	-0.593	0.371
Previous alignment	0.206 (0.085)	2.408	0.016	0.024	0.393
Interaction effects					
Magic x Disagreement valence	0.667 (0.136)	4.919	8.69e-07	0.406	0.947
Turbulence x Disagreement valence	-0.761 (0.132)	-5.764	8.21e-09	-1.049	-0.505

Note: Columns show parameter estimates, standard errors, z-values, p-values, lower and upper bounds of the bootstrapped 95% Confidence Intervals (bci). Parameter estimates standard error and their associated p-values were obtained using Wald method (Bates et al., 2015). "x" denotes interaction terms.

Figure S2 | Parameter estimates, with bootstrapped 95% confidence intervals, from the logistic model of JoA alignment ($\mathcal{M}_3$), on (i) the sample including the participants who never aligned their agency ratings to the advisor's feedback (in green), and (ii) the sample excluding these participants (in black).



10.5 Post hoc analysis about *Turb*Disagreement valence and Magic*Disagreement valence* interaction terms of $\mathcal{M}_3$

With regards to Hypothesis #3 (see section 3.3), we found that, as expected, the effect of Disagreement valence on JoA alignment was weaker in the Turbulence condition than in the other game conditions (see section 4.3). Post-hoc tests on the $\mathcal{M}_3$ model further showed several interesting features. First, in the Turbulence condition, participants were less likely to align their JoA #2 when advisors provided positive disagreements than when they provided negative disagreements (β value for the difference between positive and negative disagreements conditions = -0.457, Wald test z-score = -1.997, p = 0.023; see *Figure 4B*). Interestingly, an opposite effect was observed in both the Magic and Control conditions (β value for the difference between positive and negative disagreements conditions in the Magic game condition = 0.980, Wald test z-score = 5.107, p = 2e-07; β value for the difference between positive and negative disagreements conditions in the Control game condition = 0.391, Wald test z-score = 2.190, p = 0.014; see *Figure 4B*). Second, when disagreements were positively signed, participants were less likely to align their JoA #2 in the Turbulence condition than in the other two game conditions (β value for the difference between Turbulence and Control conditions = -0.4108061, Wald test z-score = -2.625, p = 0.004; β value for the difference between Turbulence and Magic conditions = -0.918,

Wald test z-score = -4.818, *p* = 7.234e-07; see *Figure 4B*). In contrast, when disagreements were negatively signed, participants were more likely to align their JoA #2 in the Turbulence condition than in the other conditions (β value for the difference between Turbulence and Control conditions = 0.438, Wald test z-score = 2.858, *p* = 0.002; β value for the difference between Turbulence and Magic conditions = -0.519, Wald test z-score = 3.070, *p* = 0.001; see *Figure 4B*).

We also found that the effect of Disagreement valence on JoA alignment was higher in the Magic condition than in other conditions (see section 4.3). Post-hoc tests on the $\mathcal{M}_3$ model indeed showed that participants were more likely to align their JoA #2 in the Magic condition than in other game conditions when disagreements were positive (β value for the difference between Magic and Control conditions = 0.507, Z-score = 2.751, *p* = 0.003; β value for the difference between Magic and Turbulence conditions = 0.918, Z-score = 4.818, *p* = 7.234e-07; see *Figure 4B*) than when they were negative (β value for the difference between Magic and Turbulence conditions = - 0.519, Wald test z-score = - 3.070, *p* = 0.001; see *Figure 4B*).

10.6 independent version of $\mathcal{M}_3$ including past agreement with the same advisor ($\mathcal{M}_{3b}$)

Table S5 | Logistic mixed model of JoAs alignment, including past agreement instead of past JoA aligment during past interaction with the same advisor ($\mathcal{M}_{3b}$ Nobs = 4172; Nsubj = 29).

Regression coefficients of main and interaction effects. The effect of primary interest (past agreement) is highlighted in bold.

	Estimate (SE)	***z*-value**	***p*-value**	**2.5%**	**97.5%**
Main effects					
Intercept	-0.881 (0.169)	-5.220	1.79e-07	- 1.237	- 0.570
Turbulence	-0.064 (0.070)	-0.908	0.364	- 0.205	0.081
Magic	0.141 (0.078)	1.830	0.067	- 0.006	0.287
Performance (Hit rate)	-0.194 (0.056)	-3.448	5.65e-4	- 0.314	- 0.093
JoAs #1	-0.007 (0.056)	-0.129	0.898	- 0.119	0.093

Disagreement strength	0.597 (0.156)	3.822	1.32e-4	0.292	0.906
Disagreement valence	0.301 (0.149)	2.013	0.044	0.033	0.618
Offline mean JoA	-0.335 (0.214)	-1.569	0.117	- 0.733	0.093
Past agreement	**0.119 (0.080)**	**1.481**	**0.139**	**- 0.045**	**0.280**
Interaction effects					
Magic x Disagreement valence	0.676 (0.138)	4.884	1.04e-06	0.413	0.945
Turbulence x Disagreement valence	-0.757 (0.134)	-5.663	1.49e-08	- 1.014	- 0.497

Note: Columns show parameter estimates, standard errors, z-values, p-values, lower and upper bounds of the bootstrapped 95% Confidence Intervals (bci). Parameter estimates standard error and their associated p-values were obtained using Wald method (Bates et al., 2015). "x" denotes interaction terms.

Past agreement represents the advisor's past agreement with the participant during their previous interaction, and was coded as *Past agreement* = 0.5, *Past disagreement* = -0.5.

We computed VIF for all the regressors. We found an average VIF of 1.516 s.d. ± 0.434 (median: 1.332, max: 2.348), suggesting that multicollinearity between regressors was not an issue in our model (Kock & Lynn, 2012).

10.7 Extensions of $\mathcal{M}_3$ investigating the impact of advisors' facial features on JoAs alignment

Table S6 | Logistic mixed model of JoAs alignment, including the variable representing the dimension along which the facial traits of the advisor vary (dominance versus trustworthiness dimensions) ($\mathcal{M}_{3c;}$ Nobs = 4172; Nsubj = 29).

Regression coefficients of main and interaction effects. The effect of primary interest (Facial trait dimension) is highlighted in bold.

	Estimate (SE)	***z*-value**	***p*-value**	**2.5%**	**97.5%**
Main effects					
Intercept	-0.934 (0.188)	-4.965	6.86e-07	- 1.329	- 0.573
Turbulence	-0.041 (0.071)	-0.583	0.560	- 0.183	0.099
Magic	0.113 (0.078)	1.454	0.146	- 0.046	0.265
Performance (Hit rate)	-0.171 (0.057)	-2.975	0.003	- 0.276	- 0.062
JoAs #1	-0.001 (0.057)	-0.011	0.991	- 0.116	0.119
Disagreement strength	0.603 (0.158)	3.818	1.35e-4	0.290	0.937
Disagreement valence	0.302 (0.152)	1.981	0.048	- 0.001	0.607
Offline mean JoA	-0.329 (0.210)	-1.568	0.117	- 0.747	0.124
Previous alignment	0.137 (0.087)	1.572	0.116	- 0.039	0.310
Facial trait dimension (dummy coded)	**0.063 (0.140)**	**0.451**	**0.651**	**- 0.190**	**0.324**
Interaction effects					

Magic x Disagreement valence	0.678 (0.137)	4.951	7.39e-07	0.409	0.968
Turbulence x Disagreement valence	- 0.765 (0.133)	- 5.734	9.81e-09	- 1.035	- 0.522

Note: Columns show parameter estimates, standard errors, z-values, p-values, lower and upper bounds of the bootstrapped 95% Confidence Intervals (bci). Parameter estimates standard error and their associated p-values were obtained using Wald method (Bates et al., 2015). "x" denotes interaction terms.

Trials in which the participants interacted with advisors varying on either the dominant or trustworthiness dimension were divided into two distinct “social” blocks. *Facial trait dimension* represents each of these blocks, and was dummy coded as *Dominance dimension* = 1 (corresponding to trials in which the advisor's facial trait varied along the dominance dimension), *Trustworthiness dimension* = 0 (corresponding to trials in which the advisor's facial trait varied along the trustworthiness dimension).

We computed VIF for all the regressors. We found an average VIF of 1.462 s.d. ± 0.421 (median: 1.228, max: 2.349), suggesting that multicollinearity between regressors was not an issue in our model (Kock & Lynn, 2012).

Table S7 | Logistic mixed model of JoAs alignment, including trustworthy vs untrustworthy advisors' facial traits ($\mathcal{M}_{3d}$; Nobs = 2086; Nsubj = 29).

Regression coefficients of main and interaction effects. The effect of primary interest (Trustworthiness) is highlighted in bold.

	Estimate (SE)	***z*-value**	***p*-value**	**2.5%**	**97.5%**
Main effects					
Intercept	- 0.933 (0.188)	- 4.958	7.14e-07	- 1.316	- 0.574
Turbulence	- 0.091 (0.097)	- 0.932	0.35141	- 0.292	0.112
Magic	0.201 (0.110)	1.826	0.06787	- 0.008	0.444
Performance (Hit rate)	- 0.220 (0.090)	-2.446	0.01445	- 0.397	- 0.046
JoAs #1	0.0123 (0.080)	0.153	0.87841	- 0.147	0.182
Disagreement strength	0.546 (0.177)	3.095	0.00197	0.203	0.907
Disagreement valence	0.404 (0.201)	2.006	0.04486	0.010	0.801
Offline mean JoA	- 0.362 (0.223)	- 1.625	0.10408	- 0.846	0.067
Previous alignment	- 0.051 (0.125)	- 0.408	0.68339	- 0.319	0.197
Trustworthiness	**- 0.023 (0.108)**	**- 0.216**	**0.829**	**- 0.244**	**0.176**
Interaction effects					
Magic x Disagreement valence	0.647 (0.221)	2.937	0.003	0.227	1.084
Turbulence x Disagreement valence	- 0.761 (0.188)	- 4.047	5.20e-05	- 1.134	- 0.405

Note: Columns show parameter estimates, standard errors, z-values, p-values, lower and upper bounds of the bootstrapped 95% Confidence Intervals (bci). Parameter estimates standard error and their associated p-values were obtained using Wald method (Bates et al., 2015). "x" denotes interaction terms.

Trustworthiness represents the facial traits of the advisors that vary according to the trustworthiness dimension in one of the "social" blocks, and was coded as *Trustworthy* = 0.5, un-*trustworthy* = -0.5.

We computed VIF for all the regressors. We found an average VIF of 1.435 s.d. ± 0.401 (median: 1.187, max: 2.224), suggesting that multicollinearity between regressors was not an issue in our model (Kock & Lynn, 2012).

Table S8 | Logistic mixed model of JoAs alignment, including dominant vs non dominant advisors' facial traits ($\mathcal{M}_{3e;}$ Nobs = 2086; Nsubj = 29).

Regression coefficients of main and interaction effects. The effect of primary interest (Dominance) is highlighted in bold.

	Estimate (SE)	***z*-value**	***p*-value**	**2.5%**	**97.5%**
Main effects					
Intercept	- 0.956 (0.204)	- 4.683	2.83e-06	- 1.395	- 0.582
Turbulence	- 0.024 (0.111)	- 0.219	0.827	- 0.229	0.208
Magic	0.030 (0.123)	0.248	0.804	- 0.215	0.264
Performance (Hit rate)	- 0.095 (0.0815)	- 1.161	0.246	- 0.254	0.079
JoAs #1	- 0.008 (0.0830)	- 0.092	0.926	- 0.165	0.160
Disagreement strength	0.711 (0.196)	3.631	2.83e-4	0.340	1.109
Disagreement valence	0.208 (0.172)	1.208	0.227	- 0.133	0.556
Offline mean JoA	- 0.303 (0.250)	- 1.212	0.226	- 0.770	0.194
Previous alignment	0.312 (0.132)	2.370	0.018	0.052	0.601
Dominance	**- 0.035 (0.112)**	**- 0.312**	**0.755**	**- 0.263**	**0.174**
Interaction effects					
Magic x Disagreement valence	0.715 (0.186)	3.849	1.19e-4	0.326	1.098
Turbulence x Disagreement valence	- 0.790 (0.201)	- 3.935	8.33e-05	- 1.217	- 0.391

Note: Columns show parameter estimates, standard errors, z-values, p-values, lower and upper bounds of the bootstrapped 95% Confidence Intervals (bci). Parameter estimates standard error and their associated p-values were obtained using Wald method (Bates et al., 2015). "x" denotes interaction terms.

Dominance represents the facial traits of the advisor that vary according to the dominance dimension in one of the "social" blocks, and was coded as *Dominant* = 0.5, *Non-dominant* = -0.5.

We computed VIF for all the regressors. We found an average VIF of 1.495 s.d. ± 0.446276 (median: 1.280, max: 2.544), suggesting that multicollinearity between regressors was not an issue in our model (Kock & Lynn, 2012).