Dear Author,

Any queries or remarks that have arisen during the processing of your manuscript are listed below and highlighted by flags in the proof. Please check your proof carefully and mark all corrections at the appropriate place in the proof (e.g., by using on-screen annotation in the PDF file) or compile them in a separate list.

For correction or revision of any artwork, please consult http://www.elsevier.com/artworkinstructions.

**Articles in Special Issues:** Please ensure that the words 'this issue' are added (in the list and text) to any references to other articles in this Special Issue.

| Uncited references: References that occur in the reference list but not in the text – please position each reference in the text or delete it from the list. |
| Missing references: References listed below were noted in the text but are missing from the reference list – please make the list complete or remove the references from the text. |

<table>
<thead>
<tr>
<th>Location in article</th>
<th>Query / remark</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>Please check and approve the edit of affiliations.</td>
</tr>
<tr>
<td>Q2</td>
<td>Please check the placement of Fig. 7.</td>
</tr>
</tbody>
</table>

**Electronic file usage**

Sometimes we are unable to process the electronic file of your article and/or artwork. If this is the case, we have proceeded by:

- [ ] Scanning (parts of) your article
- [ ] Rekeying (parts of) your article
- [ ] Scanning the artwork

Thank you for your assistance.
Probability judgments of agency: Rational or irrational?

Thomas Schmidt a,∗, Vera C. Heumüller b

a Faculty of Social Sciences, Psychology I, University of Kaiserslautern, Erwin-Schrödinger-Str. Geb. 57, D-67655 Kaiserslautern, Germany
b Department of Experimental and General Psychology, University of Giessen, Germany

Abstract

We studied how people attribute action outcomes to their own actions under conditions of uncertainty. Participants chose between left and right keypresses to produce an action effect (a corresponding left or right light), while a computer player made a simultaneous keypress decision. In each trial, a random generator determined which of the players controlled the action effect at varying probabilities, and participants then judged which player had produced it. Participants' effect control ranged from 20% to 80%, varied blockwise, and they could use trial-by-trial feedback to optimize the accuracy of their agency judgments. Participants tended to attribute action effects to themselves (agency bias), probably reflecting a rational guessing strategy of always naming the more likely player. However, participants systematically neglected information favoring the computer player as the agent, even under conditions where this bias could only harm judgment accuracy. We conclude that agency biases have both rational and irrational components.

© 2010 Published by Elsevier Inc.

1. Introduction

Our sense of agency is vital to our ability to act in a coherent and successful manner. The sense of agency may be defined as the subjective impression that one has caused the consequences of one's own actions, leading to a self-experience as an autonomous, freely acting individual (Frith, 2002; Metcalfe & Greene, 2007). In a wider meaning, adopted in this paper, the sense of agency can be viewed as an attribution of action effects to one's own actions, with or without a distinct subjective impression of agency. Wegener and Wheatley (1999; Wegner, 2002) stress the role of intentions for agency attributions: in their view, people perceive themselves as agents if their intention to act occurs before the action, is consistent with it, and if its consequences are not easily attributable to outside causes.

It is instructive to examine situations where agency is misattributed. For example, obsessive–compulsive disorder is characterized by massive misperceptions of control (Moulding & Kyrios, 2006): on the one hand, patients experience a loss in action efficacy, e.g., when repeatedly checking that the stove is switched off; on the other hand, they engage in ever-repeating rituals that they believe will restore their control. Patients with obsessive–compulsive disorder overestimate the degree of control they have over sequences of purported action effects that are actually completely preprogrammed (Reuven-Magril, Dar, & Liberman, 2008). In contrast, underestimation of action control can occur in table-turning games where control is jointly exerted by all players but falsely attributed to an outside agent, even an otherworldly spirit.

Even in normal observers, the sense of control can be strongly distorted. In a classical study, Alloy and Abramson (1979) reported that participants overestimate the contingency between their own keypress actions and the subsequent lighting up of a response light, whereas patients diagnosed with depression were less prone to such control illusions.

∗ Corresponding author. Fax: +49 631 2053910.
E-mail address: thomas.schmidt@sowi.uni-kl.de (T. Schmidt).

1053-8100/$ - see front matter © 2010 Published by Elsevier Inc.
doi:10.1016/j.concog.2010.01.004

Please cite this article in press as: Schmidt, T., & Heumüller, V. C. Probability judgments of agency: Rational or irrational? Consciousness and Cognition (2010), doi:10.1016/j.concog.2010.01.004
Linser and Goschke (2007) showed that the subliminal priming of action effects enhanced perceived control, even if the effects were in fact unrelated to the participants’ actions (also see Aarts, 2007; Aarts, Custors, & Wegner, 2005; Kiesel et al., 2006). Repp and Knoblich (2007) asked participants to synchronize repetitive tapping movements to a rhythmical tone signal, and to detect the transition between a phase where the tone signal was computer-generated and a phase where it was controlled by the participants’ finger taps (or vice versa). Most participants had trouble finding this transition point, revealing a robust bias to interpret tone signals as effects of their own actions. Similarly, Wegner and Wheatley (1999) showed that participants can be fooled to believe that they have moved a mouse pointer when in fact a confederate has. These studies allow the conclusion that people have an inclination to attribute action effects to their own actions. This is what we call an agency bias.

The examples of agency biases described so far may all be viewed as a rational response of a system evolved to monitor whether self-produced actions led to the desired consequences. Indeed, Metcalfe and Greene (2007) demonstrated that participants are quite adept at separating their own amount of action control (e.g., in moving a joystick to control a cursor) from external perturbations of the movements (e.g., random noise added to the cursor position). However, other studies indicate that the bias to attribute action effects to oneself can be highly irrational or even bizarre. In a study by Wegner, Sparrow, and Winerman (2004), participants watched themselves in a mirror such that a confederate’s hands appeared in the position of their own hands. The confederate was then instructed to move the hands. When the participants could hear these instructions prior to the action, they felt themselves to be causing the other person’s actions. Pronin, Wegner, McCarthy, and Rodriguez (2006) had participants perform a voodoo ritual that was supposed to harm a person, after which the purported victim of the curse complained about headaches. Participants had a stronger sense of actually having harmed the victim when they had been induced to harbor negative attitudes about the victim, as compared to participants with a neutral attitude.

In this paper, we define “agency” as an attribution process of an action effect to one’s own action, making no claims about the “feeling” or “sense” of agency as a private experience. In this form, the definition can be extended to attributions of other people’s action effects to their respective actions. The purpose of our experiments was to disentangle rational and irrational components in the agency bias. A rational component would mean that people tend to attribute agency to themselves because they are in fact most likely to having produced it. An irrational component would be to attribute agency to oneself even under neutral probabilities, or in situations where the evidence rather suggests a different agent. Furthermore, we wanted to see whether agency biases only occur for self-generated actions, or whether they are also applied to salient other agents.

2. Experiment 1

In each trial of Experiment 1, the participant freely chose to perform a left or right keypress to produce an action effect, which was a left or right light spatially corresponding to the chosen keypress. At the same time, an invisible computer player also chose to press a (virtual) key (equiprobably left or right), and a random generator decided whether the participant or the computer player was the agent in this trial, i.e., actually controlled the action effect. After each trial, participants had to guess from the observed effect whether they themselves or the computer player had been the agent, and received immediate feedback whether they had guessed correctly. The contingency of the participant’s effect control varied from block to block, with contingencies between 20% and 80%. The order of blocks with different contingencies was random, such that the participant had to infer the contingencies anew in each block. We expected that when agency was uncertain, participants would be biased to attribute agency to themselves rather than to the invisible computer player.

2.1. Methods

2.1.1. Participants

Eight students from the University of Giessen (average age 23.5 years, four male) volunteered for course credits or for a payment of € 8/h. Their vision was normal or appropriately corrected. All of them gave informed consent.

2.1.2. Stimuli and procedure

The experiment was controlled by a 300-MHz personal computer driving a 14” VGA color monitor (640 by 480 pixel) in synchrony with the monitor retrace rate of 60 Hz. The monitor was viewed from a distance of approximately 1 m in a dimly lit room.

The procedure is shown in Fig. 1. Each trial started with the appearance of two circle-shaped placeholders (diameter 1.29”) 2.10° to the left and right of a fixation point. Participants performed a free-choice keypress action (“1” or “3” on the numerical keypad). Immediately after the keypress, either the left or the right placeholder turned red for 200 ms (action effect), either in response to the participant’s keypress or to the decision of an invisible computer player. The computer player was simulated by a random generator which decided on “left” or “right” keypresses with equal probability. Another random generator decided whether the action effect was based on the participant’s or the computer player’s decision. Following the action effect, a mouse cursor appeared in the center of the screen, and participants had to click on the name of the...
player they believed had controlled the action effect (either “Rudi”, the computer player’s name, or “Ich” (“Me”). The two
names appeared above and below the fixation point, respectively, and switched position randomly from trial to trial. When
the agent had been guessed correctly, a 1000-Hz, 200-ms tone was sounded; otherwise, there was a 100-Hz, 200-ms tone.

Participants performed three sessions, each consisting of seven blocks of 100 trials. The contingency at which participants
actually controlled the action effect varied from block to block and ranged from 20% to 80% in 10-% steps. The order of these
blocks was randomized, so that participants had to pick up the contingency anew in each block. In addition to the auditory
feedback after each trial, participants received summary feedback about their percentage of correct responses after each
block. All participants were debriefed after the end of the experiment.

Note that this task is trickier than may be apparent from the straightforward design. The participant can only be sure
about the actual agent in those cases where participant and computer player decide on different keys and it is the computer
player who controls the action effect. Agency judgments in these trials are trivial because the observed action effect is inconsis-
tent with the participant’s keypress and must therefore be attributed to the computer player. In all other cases, participants
have to consider that the effect might have been controlled by the computer player even though the effect is consistent with their own action. In principle, therefore, there are two sources of information about the contingency of
the participant’s effect control in a given block: the auditory feedback about the correctness of their agency attributions
in each trial, and the overall percentage of trivial trials encountered during the block. We carefully instructed participants
until we could be certain that they had fully understood the probability structure of the experiment. In particular, we explic-

1 Names were given to the computer players to emphasize to the participants that the computer player performed exactly the same function as they did,
freely choosing keypress actions, just as if they were real persons. It was, however, clear that “Rudi” was a computer routine and not a human player working by remote.
It might be expected that this strong agency bias would be quite detrimental to judgment accuracy, but this was mostly with each contingency level, in 67% of the nontrivial trials, they judged themselves to be in control in 84% (Fig. 2a). This agency bias was significant at control given only the nontrivial trials is therefore

\[ p_{\text{trivial}}(c) = \frac{1 - c}{2}; \]  

\[ p_{\text{participant}}(c) = \frac{c}{1 - (1 - c)/2}; \]  

\[ p_{\text{computer}}(c) = \frac{(1 - c)/2}{1 - (1 - c)/2}; \]  

i.e., half of those trials where the participant is not in control (assuming that the computer’s keypress decisions are the same as the participant’s in half of the trials). Trivial trials were excluded from analysis. The probability that the participant is in control given only the nontrivial trials is therefore

\[ p_{\text{participant}}(c) = \frac{c}{1 - (1 - c)/2}; \]  

\[ p_{\text{computer}}(c) = \frac{(1 - c)/2}{1 - (1 - c)/2}; \]  

2.1.3. Stochastic analysis

Trials were considered trivial when the action effect was obviously controlled by the computer player, which occurred in those trials where the computer player controlled the effect and had decided on a different key than the participant. Let \( c \) be the contingency at which the participant controls the effect. Then, the proportion of trivial trials is expected to be

\[ p_{\text{trivial}}(c) = \left( 1 - c \right)/2; \]  

\[ p_{\text{participant}}(c) = \frac{c}{1 - (1 - c)/2}; \]  

\[ p_{\text{computer}}(c) = \frac{(1 - c)/2}{1 - (1 - c)/2}; \]  

i.e., the proportion \( c \) out of all nontrivial trials. Likewise, the probability that the computer player is in control given only the nontrivial trials is

\[ p_{\text{participant}}(c) = \frac{c}{1 - (1 - c)/2}; \]  

\[ p_{\text{computer}}(c) = \frac{(1 - c)/2}{1 - (1 - c)/2}; \]  

i.e., the proportion \( 1 - c/2 \) of trials where the participant is not in control but decides on the same key as the computer player, out of all nontrivial trials. Note that \( p_{\text{participant}} > p_{\text{computer}} \) in most conditions; actually, it is only in the 20% and 30% conditions that the computer player is more likely than the participant to control the action effect. This is a consequence of the fact that only the trivial trials allow for unequivocal agency attributions to the computer player, whereas agency attributions to the participant can never be made with certainty. As a result, trials where the participant is the actual agent are overrepresented in those trials where it is necessary to guess the agent, i.e., in the nontrivial trials.

Eqs. (2) and (3) give the expected percentage of correct responses when participants attribute agency exclusively to themselves or exclusively to the computer player, respectively. Therefore, the optimal guessing strategy is to always indicate the player as the agent who is more likely to control the action effect, which guarantees that the percentage of correct responses is \( p_{\text{participant}} \) whenever \( p_{\text{participant}} > p_{\text{computer}} \), and \( p_{\text{computer}} \) whenever \( p_{\text{participant}} < p_{\text{computer}} \). In sum, it is optimal to always indicate oneself as the agent except when the trial is trivial or at contingency conditions of 30% or lower. We also considered that participants might engage in a probability-matching strategy. Under such a strategy, when participants believe they are in control with probability \( \beta \), they indicate themselves as the agent in a proportion \( \beta \) of trials. Assuming that participants’ estimate of \( \beta \) is correct, the percentage of correct responses under this strategy would be

\[ p_{\text{probmatch}}(\beta) = \beta^2 + (1 - \beta)^2; \]  

i.e., a proportion \( 1 - \beta \) of those trials where the participant was not in control, and a proportion \( \beta \) of those trials where the participant was in control. Assuming that participants are able to correctly estimate that \( \hat{\beta} = p_{\text{participant}} \) in nontrivial trials, the expected percentage of correct responses would be

\[ p_{\text{probmatch}}(\hat{\beta}) = \left[ \frac{c}{1 - (1 - c)/2} \right]^2 + \left[ \frac{(1 - c)/2}{1 - (1 - c)/2} \right]^2. \]  

2.1.4. Statistical methods

Only nontrivial trials were considered. Data from one participant had to be dismissed because she always indicated herself as the agent irrespective of contingency in nontrivial trials, and even in a high proportion of trivial trials, suggesting that she had not followed task instructions. We analyzed the discrepancy of agency judgments and response accuracies from various model predictions by binomial \( z \) tests, pooling responses across participants. In addition, we fit the data of individual participants to the different models using a \( \chi^2 \) statistic, and then combined participants by summing up their \( \chi^2 \) values and degrees of freedom (Fleiss, 1981). Note that both methods focus on single-case analysis. When comparing experiments, additional analyses of variance (ANOVM, Huynh–Feldt corrected) were performed on the arcsine-transformed percentage scores to render the binomial data compatible with ANOVA requirements.

2.2. Results and discussion

Participants indicated themselves as the agent systematically too often: for example, when they were actually in control in 67% of the nontrivial trials, they judged themselves to be in control in 84% (Fig. 2a). This agency bias was significant at each contingency level, \( 3.90 < z < 16.60, \) all \( p < .0001 \). Consequently, comparing the percentage of “agent is me” judgments with \( p_{\text{participant}} \) in a goodness-of-fit test yielded a large discrepancy, \( \chi^2(42) = 657.32, \) \( p < .0001 \).

It might be expected that this strong agency bias would be quite detrimental to judgment accuracy, but this was mostly not the case (Fig. 2b). Remember that the optimal strategy is to indicate oneself as the agent whenever \( p_{\text{participant}} > p_{\text{computer}} \), and the computer player as the agent whenever \( p_{\text{participant}} < p_{\text{computer}} \), which is the case only at contingency levels of 20% and 30%.
There are clear individual differences in agency attributions (Fig. 3), probably reflecting different guessing strategies. In all participants, the percentage of “agent is me” judgments increased with contingency, and most participants showed a clear bias to attribute agency to themselves (with the possible exception of participants C and F). Only two participants (A and F) systematically named the computer player in the low-contingency conditions. Participants A and B used a strategy of always naming themselves, which was followed by the computer player in the 20% condition. Here, performance was better than under a strategy where one would always indicate oneself as the agent (always-me strategy), $z = 14.51, p < .0001$, but still worse than the optimal strategy, $z = -10.78, p < .0001$. Overall, performance was closer to the optimal strategy, $\chi^2(42) = 101.59$, than to the always-me strategy, $\chi^2(42) = 209.97$; actually, most of the data fell between the optimal strategy and simple probability matching, which would yield $\chi^2(42) = 84.67$. However, all these models still departed significantly from the data, all $p < .0001$, indicating that neither of the described strategies was followed exclusively.

Because contingencies had to be inferred anew in each block, we analyzed how agency judgments might change over the course of a block. We found that agency judgments stabilized quickly and did not change after the first quarter of trials in each block. The only exception was the 20% condition, where agency was attributed to the self more often (by 10–20 percentage points) in the first quarter of trials than in the remaining block. An ANOVA with factors of session, contingency and block quarter revealed (besides the strong main effect of contingency) an interaction effect of contingency and quarter, $F(18, 90) = 3.32, p = .0045$, but no further effects involving the quarter or session factor. This pattern indicates that at the outset of a new low-contingency block, participants gradually learned to name the computer player more often, an effect not apparent in the remaining contingency conditions. Therefore, even though the agency bias might be stronger at the outset of a new low-contingency block (probably due to a lack of information), it is never substantially remedied as contingency information is being picked up.

In sum, Experiment 1 confirms a strong agency bias: participants tend to attribute action effects to their own actions rather than to those of an invisible computer player. In terms of judgment accuracy, however, this bias appears largely rational and is detrimental to performance in the 20% and 30% conditions only. However, even though participants actually tended to favor the computer player in those conditions, that switch in agency attributions was not large enough to be called optimal. In other words, the agency bias enabled participants to exploit information favoring themselves as the agent, but the participants failed to fully exploit information favoring the computer player. This asymmetry suggests that there might be a more rational component to the agency bias.

One interesting question is whether agency biases only arise when agency is attributed to oneself, or whether agency is also misattributed to a salient other person. Of course, this problem requires extending our former definition of “agency” to include actions of other people. Under that definition, perceiving another person as an “agent” consists in an attribution of action effects to that person’s actions. We are aware that this definition blurs the line between agency and causality judgments.

2 Of course, this problem requires extending our former definition of “agency” to include actions of other people. Under that definition, perceiving another person as an “agent” consists in an attribution of action effects to that person’s actions. We are aware that this definition blurs the line between agency and causality judgments.
will we tend to overestimate the degree to which that person is in control of action effects, just as we overestimate control in
ourselves? In Experiment 2, we designed a situation where participants only observed a computer player (focal player) making
choices but did not act themselves, in a setting otherwise identical to Experiment 1. Again, a random generator decided whether
action effects were determined by the overt choices of this focal player or the covert choice of an invisible second computer
player (other player). If agency is misattributed on the basis of which player is most salient, we would expect that participants
would be biased to attribute agency to the focal player rather than to the invisible other player. However, just as in Experiment
1, such a bias might still reflect a rational strategy of systematically choosing the more likely player. In session 2, the actions of
the formerly invisible other player were revealed, so that the participant was fully informed about both players’ actions. The
question was whether a bias towards the formerly salient focal player would persist even though there was no rational basis
for still favoring that player. Clearly, any persisting agency bias would now be irrational because it could only harm judgment
accuracy.

3. Experiment 2

Experiment 2 was similar to Experiment 1, except that the participants now observed two computer players and did not
perform any actions themselves. There were three sessions. In each trial of sessions 1 and 3, the participant observed the
overtkeypress decision of one computer player (focal player), while a second computer player (other player) made a keypress
decision covertly. A random generator decided which player was in control of the action effect, and the participant had to
guess this player. Again, the contingency of the focal player’s effect control varied from block to block, with contingencies
between .2 and .8.

3.1. Methods

3.1.1. Participants

Eight new students from the University of Giessen (average age 23.1 years, three male) volunteered for course credits or
for a payment of €8/h. Their vision was normal or appropriately corrected. All of them gave informed consent.

3.1.2. Stimuli and procedure

The experimental setup was the same as in Experiment 1. The procedure is shown in Fig. 4. Again, each trial started with
the appearance of two circle-shaped placeholders (same as in Experiment 1), this time together with square-shaped place-
holders (0.53 × 0.53°, 1.49° from the fixation dot) symbolizing the action decisions of the two computer players. The focal
player’s action decisions were always displayed below fixation, and those of the other player always above fixation. In

Fig. 3. Proportion of “agent is me” judgments in single participants (a–g) in Experiment 1. Standard errors are between trials.
sessions 1 and 3, only the two placeholders symbolizing the focal player’s decision were shown. Participants started the trial by pressing the space button. The keypress decision of the focal player was revealed by letting the lower left or right of the square placeholders turn red for 500 ms. In session 2 only, this was followed by the keypress decision of the other player by letting the upper left or right of the square placeholders turn red for 500 ms. After a delay of 500 ms, the action effect was presented: either the left or right circular placeholder turned red for 200 ms, either in response to the focal player’s or the other player’s decision. Both computer players were simulated by random generators which decided on “left” or “right” responses with equal probability. Another random generator decided which player controlled the action effect. Following the action effect, a mouse cursor appeared in the center of the screen, and participants had to click on the name of the player they believed had controlled the action effect (either “Paul”, the focal player’s name, or “Rudi”, the other player’s name). The two names appeared above and below the fixation point, respectively, and switched position randomly from trial to trial. When the choice was correct, a 1000-Hz, 200-ms tone was sounded; otherwise, there was a 100-Hz, 200-ms tone.

Each session consisted of seven blocks of 100 trials. The contingency at which the focal player controlled the action effect varied from block to block and ranged from 20% to 80% in 10-% steps. The order of these blocks was randomized, so that participants had to infer the contingency anew in each block (using the immediate feedback and the proportion of trivial trials in a block). Again, participants were informed beforehand that seven contingency levels would be used in random sequence, one for each block, and what those contingencies would be. In addition to the auditory feedback after each trial, participants received summary feedback about their percentage of correct responses after each block. All participants were debriefed after the end of the experiment.

3.1.3. Stochastic analysis

In sessions 1 and 3, trials were considered trivial when the action effect was obviously controlled by the other player, which again occurred with probability $p_{\text{trivial}}(c) = (1 - c)/2$, $c$ now being the contingency at which the focal player controlled the action effect. The probabilities that the focal or the other player is in control given only the nontrivial trials is obtained by letting $p_{\text{participant}} = p_{\text{focal}}$ and $p_{\text{computer}} = p_{\text{other}}$ in Eqs. (2) and (3). The proportion of correct responses under a probability-matching strategy is also unchanged (Eq. (4)). In session 2, trials were trivial whenever the focal and the other player differed in their choices, because this difference was always visible. In that session, the probabilities that the focal or the other player are in control given only the nontrivial trials are simply $c$ and $(1 - c)$, respectively, and the proportion of correct responses expected under a probability-matching strategy becomes $p_{\text{probmatch}}(c) = c^2 + (1 - c)^2$. 

Fig. 4. Methods used in Experiment 2. (a) In Sessions 1 and 3, participants observed the action decision of a focal player (indicated by the lighting up of the left or right of the two small squares), followed by the action effect after a 500-ms delay. Then, participants performed an agency judgment (via mouse cursor) indicating whether the effect had been caused by the focal or another player whose actions were determined by a random generator. (b) In Session 2, both players’ actions were visible. (c) The contingency at which the focal player controlled the action effect varied between 20% and 80%. Contingency was varied between blocks of 100 trials, with the sequence of blocks determined randomly.
3.2. Results and discussion

Results were pooled across sessions 1 and 3 because ANOVAs of agency judgments and response accuracies yielded no significant main effects or interactions for a “session” factor.

3.2.1. Sessions 1 and 3

Participants still indicated the focal player to be in control too often (Fig. 5a), but only at contingencies above 50% (60%: \( z = 8.40; 70\%: z = 9.87; 80\%: z = 8.63, \text{all } p < .0001 \)). There was significant underestimation of the focal player’s agency for contingencies below 50% (20%: \( z = -7.02, p = .0002 \); 30%: \( z = -5.74, p = .0020 \); all other \( p > .2 \)). In comparison with Experiment 1, the agency bias was diminished, \( F(1, 13) = 6.10, p = .0282 \). All in all, however, comparing the percentage of “agent is focal player” judgments with \( p_{\text{local}} \) in a goodness-of-fit test still yielded a large discrepancy, \( \chi^2(48) = 473.57, p < .0001 \).

Judgment accuracy in the 20% condition was better than under a strategy where one would always indicate the focal player as the agent (always-focal strategy), \( z = 16.34, p < .0001 \), but significantly worse than the optimal strategy, \( z = -5.75, p < .0001 \) (Fig. 5b). Across all contingency conditions, performance was closer to the optimal strategy, \( \chi^2(48) = 100.86 \), than to the always-focal strategy, \( \chi^2(48) = 274.74 \); again, most of the data points fell between the optimal strategy and simple probability matching, \( \chi^2(48) = 73.80 \). However, all these models still departed significantly from the data, \( \text{all } p < .001 \), indicating that neither of the described strategies was followed exclusively.

3.2.2. Session 2

Performance in Session 2 was of special interest: In this condition, participants were fully informed about both players’ actions, and any remaining bias towards the focal player would be detrimental to judgment accuracy. Surprisingly, participants still indicated the focal player to be in control too often (Fig. 6a) at most contingency levels (40%: \( z = 5.26, p = .0042 \); 50%: \( z = 1.46, p = .2334 \); 60%: \( z = 7.43, p = .0001 \); 70%: \( z = 6.71, p = .0004 \); 80%: \( z = 5.48, p = .0031 \)). However, they tended to underestimate the focal player’s effect control in the low-contingency conditions (20%: \( z = -2.49, p = .1068 \); 30%: \( z = -5.04, p = .0058 \)). All in all, comparing the percentage of “agent is focal player” judgments with \( p_{\text{local}} \) in a goodness-of-fit test yielded a huge discrepancy, \( \chi^2(48) = 333.69, p < .0001 \). Participants attributed agency to the focal player slightly less often than in Session 1, \( F(1, 7) = 6.68, p = .0363 \), but about as often as in Session 2, \( F(1, 7) = 2.34, p = .1701 \).

Judgment accuracy was significantly worse than under the optimal strategy (all \( p < .0008 \) except at the 50% and 60% levels (Fig. 6b). A goodness-of-fit test concluded that a probability-matching strategy provided the best fit to the data, \( \chi^2(48) = 53.99, p = .2563 \), the optimal strategy was almost as good, \( \chi^2(48) = 56.53, p = .1864 \), and the always-focal strategy was much worse, \( \chi^2(48) = 779.69, p < .0001 \). However, the most crucial finding was that the pattern of judgment accuracies was not totally asymmetric: the lowest accuracy was obtained not in the 50% condition, but in the 40% condition. Planned comparisons revealed that accuracy in the 40% condition was significantly lower than in the comparable 60% condition, \( z = -3.28, p = .0005 \), while the 30% and 70% conditions and the 20% and 80% conditions were not reliably different from each other, \( z = -1.03 \) and \( -0.06, p = .1513 \) and .4757, respectively. Note that response accuracy in the 40% condition was low not because average agency judgments were biased, but because agency of the focal player was overestimated by some participants and underestimated by others. This pattern of response accuracy clearly differed from that obtained in Sessions 1 and 3.

![Fig. 5.](image-url) (a) Proportion of “agent is focal player” judgments in Experiment 2, Sessions 1 and 3. Only nontrivial trials are considered. The stippled line gives the probability that the focal player controlled the action effect given only the nontrivial trials (\( p_{\text{local}} \)). (b) Accuracy of agency judgments in Experiment 2, Sessions 1 and 3. Only nontrivial trials are considered. The stippled lines indicate expected performance, given only the nontrivial trials, when the participant would always indicate the focal player (\( p_{\text{local}} \)) or the other player (\( p_{\text{other}} \)) as the agent, or engage in a probability-matching strategy (\( p_{\text{probmatch}} \)). Again, the optimal strategy would be to always attribute agency to the more likely player.

Please cite this article in press as: Schmidt, T., & Heumüller, V. C. Probability judgments of agency: Rational or irrational? *Consciousness and Cognition* (2010), doi:10.1016/j.concog.2010.01.004
302 ANOVAs with factors of contingency and session showed clear interactions of both factors (Session 2 vs. Session 1: $F(6, 42) = 12.11, p < .0001$; Session 2 vs. Session 3: $F(6, 42) = 9.10, p = .0002$), but no main effect of session (see Fig. 7).

3.2.3. Strategic effects

Individual differences in guessing strategies were more pronounced than in Experiment 1. Broadly, there seem to be two major strategies. One is to match the proportion of “agent is focal player” judgments to the contingency condition, so that agency judgments become a continuous function of contingency (participants H, J, K, M, N). The other strategy seems to consist in classifying blocks into those that are predominantly controlled by the focal or the other player, respectively, and then to always name the more likely player (participants I, L, O). Note that under this strategy, blocks from similar contingency conditions may sometimes be misclassified, which would explain the conspicuous zig–zag patterns for some participants.
(e.g., participant L), which occur in spite of low standard errors. Only two participants show marked differences between sessions 1 and 3: Participant J seems to have switched from a matching to a classification strategy after the first session, and participant I became more conservative in attributing agency to himself.

We again analyzed how agency judgments might change during each block. We found that agency judgments stabilized quickly and did not change after the first quarter of trials in each block. Similar to Experiment 1, there was a tendency in the low-contingency conditions to attribute agency to the focal player more often (by 10–20 percentage points) in the first quarter of trials than in the remaining block. ANOVAs with factors of contingency and block quarter revealed (besides the strong main effects of contingency) interaction effects of contingency and quarter in some of the three sessions of the experiment, f(18, 126) = 1.23, 1.92, and 2.58, p = .3016, .0199, and .0512 in sessions 1, 2, and 3, respectively, but no further effects involving the quarter factor. This pattern indicates that at the outset of a new low-contingency block, participants gradually learned to name the non-focal player more often. This effect was not apparent in the remaining contingency conditions. Again, this indicates that the agency bias might be stronger at the outset of a new low-contingency block but is never fully remedied during the time-course of the block.

In conclusion, Experiment 2 demonstrates that some agency bias remains even if participants do not act themselves but only observe the choices made by a computer player. The agency bias is still observable at high contingency levels but is reversed at low-contingency levels, probably because some participants adopt a strategy for always naming the player that seems more likely to them. Compared to Experiment 1, the agency bias is reduced. Surprisingly, however, a clear agency bias is observed in a situation where the participant has full information about both player’s choices so that any bias towards one or the other player is detrimental to performance.

4. General discussion

The experiments reported here show that participants are prone to agency biases in situations where the correct attribution of action effects is uncertain (Alloy & Abramson, 1979; Wegner & Wheatley, 1999). In Experiment 1, participants strongly overestimated their own degree of control over an action effect that could also have been caused by a computer player. When participants no longer acted themselves but only observed the actions of a computer player in Experiment 2, this bias was diminished but nonetheless persisted for high-contingency conditions.

How much of this bias can be attributed to a rational strategy for maximizing the accuracy of agency judgments? As it turns out, even a strong agency bias is a rational strategy for maximizing judgment accuracy in the task at hand. Although the participants showed a clear tendency to attribute agency to themselves, they achieved nearly optimal response accuracies in many contingency conditions, and never performed worse than predicted by a simple probability-matching strategy. Because most of the nontrivial trials were in fact controlled by the participants, attributing agency to oneself guarantees that most of these trials are correctly classified. This strategy is detrimental to performance only in the low-contingency (20% and 30%) conditions, where the computer player is actually more likely than the participant to control the effect. It is in exactly these conditions where the agency bias becomes a maladaptive strategy and where participants’ performance is notably worse than optimal.

However, even if the agency bias largely reflects a rational strategy for maximizing judgment accuracy, it persists in situations where it can only harm performance. A common theme in all experimental conditions was a fundamental asymmetry in how participants utilized information favoring one or the other player. In both experiments, only information favoring oneself or the focal player as the agent was efficiently used, whereas equally strong information in favor of the other player was neglected. This aspect of the agency bias is especially salient in Session 2 of Experiment 2, where an agency bias persisted even though participants had equal information about both players and even though both players’ degree of action control was perfectly symmetrical. Clearly, this bias can only harm the accuracy of agency judgments and may thus be regarded as an irrational component of the agency bias.

We observed marked individual differences in how different participants utilized the available information (Stanovich & West, 2000). Some participants were able to utilize information favoring the other (non-focal) player better than others, e.g., by mostly naming the other player as the agent in the low-contingency conditions of Experiment 2. Under the same conditions, other participants still favored themselves or the focal player as the agent. Because both over- and underestimation of agency leads to a loss in accuracy, the relationship between agency judgments and judgment accuracy is complex. One striking example is Session 2 of Experiment 2, which revealed an bias to favor the focal player over the other player even though both players’ actions were perfectly known and their control probabilities were symmetrical. Averaged across participants, a significant agency bias was observed in that session only in the high-contingency conditions (Fig. 6a), whereas departures from judgment accuracy predominantly occurred in the low-contingency conditions (Fig. 6b). This seeming discrepancy comes from the fact that under low-contingency conditions, some participants lose accuracy by overestimating agency of the focal player, while others lose accuracy by underestimating it. In contrast, under high-contingency conditions, the agency of the focal player is consistently overestimated, which guarantees that participants are close to optimal performance.

Our results extend the classical findings by Alloy and Abramson (1979) by showing that agency biases are not limited to the participant’s own actions but generalize to an agent that is only observed, even if it is just a computer player. Our results are in line with studies showing that people sometimes find it hard to tell whether purported action effects are actually due to their own actions or to those of external sources (Repp & Knoblich, 2007), or even to actions performed by another
participant (Wegner & Wheatley, 1999). Our finding that the agency bias is smaller when actions are only observed is in line with a study by Fernandez-Duque and Wifall (2007), who have shown that participants are better at judging probabilities in a gambling task if they are not playing it themselves but are only observing other players. What is surprising in our data is that the bias even persists when both players are computer players whose actions are fully known and completely comparable, and where the focal player is merely distinguished by being the only player observable in a previous session.

Our results also suggest that agency judgments are more than just judgments of causality relating actions to action effects. The contingencies between actions and effects were identical when participants were acting themselves and when they only observed the focal player’s actions, yet agency biases in Experiment 1 were markedly larger than those in Experiment 2. Obviously, making agency attributions to oneself invites agency biases more strongly than merely observing a salient other. On the other hand, the fact that watching the “actions” of two computer programs is sufficient for generating an agency bias is astonishing. It is interesting to speculate to what extent the focal computer player is perceived as a social agent even though it obvious lacks sentience, intentionality and self-awareness.

In our view, the agency bias is the natural result of the necessity to use external feedback to check whether our actions (or those of others) had the intended effects (Fernandez-Duque & Wifall, 2007). We propose that agency biases are not just limited to our own actions, but are also applied to actions of salient others. In that view, when observers are confronted with action effects that might plausibly come from different agents, they would tend to attribute agency to the most salient of these agents. In instances where one’s own actions are involved, the most salient agent is oneself, and the agency bias is maximal; on the other hand, if the agent is merely a computer player, saliency might be diminished and the agency bias reduced. As soon as the action outcome is consistent with the intended effects, illusions of vicarious control (Wegner et al., 2004) or superstitious belief in ritualistic actions (Pronin et al., 2006) can arise. The tendency to attribute agency exclusively to only one agent is probably evolutionarily adaptive: outside the realm of artificial laboratory tasks, effects consistent with the intended action are very likely to actually result from that action (Wegner & Wheatley, 1999; Metcalfe & Greene, 2007). The ultimate function of agency biases would then be to facilitate the interpretation of actions and action effects in an individualistic as well as a social context, e.g., as a cognitive foundation of self-efficacy and social learning (Bandura, 1977, 1997; Frith & Frith, 2007).

Acknowledgment

The authors thank Anna Seydell, Filipp Schmidt, and Karl R. Gegenfurtner.

References


