Can the Brain Strategically Go on Automatic Pilot? The Effect of If–Then Planning on Behavioral Flexibility

Tim van Timmeren[1,2], **John P. O'Doherty**[3], **Nadza Dzinalija**[4], and **Sanne de Wit**[1]

Abstract

■ People often have good intentions but fail to adhere to them. Implementation intentions, a form of strategic planning, can help people to close this intention–behavior gap. Their effectiveness has been proposed to depend on the mental formation of a stimulus–response association between a trigger and target behavior, thereby creating an "instant habit." If implementation intentions do indeed lead to reliance on habitual control, then this may come at the cost of reduced behavioral flexibility. Furthermore, we would expect a shift from recruitment of corticostriatal brain regions implicated in goal-directed control toward habit regions. To test these ideas, we performed a fMRI study in which participants received instrumental training supported by either implementation or goal intentions, followed by an outcome revaluation to test reliance on habitual versus goal-directed control. We found that implementation intentions led to increased efficiency early in training, as reflected by higher accuracy, faster RTs, and decreased anterior caudate engagement. However, implementation intentions did not reduce behavioral flexibility when goals changed during the test phase, nor did it affect the underlying corticostriatal pathways. In addition, this study showed that "slips of action" toward devalued outcomes are associated with reduced activity in brain regions implicated in goal-directed control (ventromedial prefrontal cortex and lateral orbitofrontal cortex) and increased activity of the fronto-parietal salience network (including the insula, dorsal anterior cingulate cortex, and SMA). In conclusion, our behavioral and neuroimaging findings suggest that strategic if–then planning does not lead to a shift from goal-directed toward habitual control. ■

INTRODUCTION

At the start of the new year, many people reflect on their future plans and form resolutions. However, they often fail to put their good intentions into practice (Sheeran & Webb, 2016). Strategic "if–then" plans, also known as implementation intentions, are an effective way to support the translation of intentions to actions. For example, instead of formulating an abstract plan such as "I want to lose weight," an implementation intention links the intended action to a specific cue or situation, for example, "If I get home, I will eat an apple," thereby enhancing the probability of success. Indeed, many studies have shown that implementation intentions support behavior change better than goal intentions that merely specify the intended action or outcome (Gollwitzer & Sheeran, 2006). In addition to increasing attention to the relevant cue, the effectiveness of if–then planning has been proposed to rely on creating a strong associative link between the stimulus (S) in the if-part ("home") and the response (R) in the then-part (eat an apple), in a manner akin to habits acquired through behavioral repetition (Dickinson, 1985; Thorndike, 1911). These mentally formed S–R associations may allow for automatic action initiation (Gollwitzer, 2014)—a process often referred to as strategic automaticity or "instant habits" (Gollwitzer, 1993, 1999, 2014).

The notion that merely using a verbal action-plan could be sufficient to form a habit is fascinating, because a central assumption in theories of habit formation is that this process critically depends on behavioral repetition. Support for the idea that implementation intentions accelerate habit formation comes from research showing that they increase (self-reported) automaticity (Orbell & Verplanken, 2010; Parks-Stamm, Gollwitzer, & Oettingen, 2007; Brandstätter, Lengfelder, & Gollwitzer, 2001). Therefore, implementation intentions lead to benefits in terms of efficient goal attainment (Gollwitzer, 2014; Gollwitzer & Sheeran, 2006). However, habits developed through behavioral repetition also come at a cost, namely, decreased behavioral flexibility (Dickinson, 1985). The question arises, therefore, if the use of implementation intentions also leads decreased flexibility when goals change. This can be investigated using the outcome-devaluation test, an experimental paradigm originally used in rats (Adams & Dickinson, 1981) and

[1]University of Amsterdam, The Netherlands, [2]Utrecht University, The Netherlands, [3]California Institute of Technology, Pasadena, [4]Amsterdam UMC, Location VUmc, The Netherlands

Journal of Cognitive Neuroscience 35:6, pp. 957–975
https://doi.org/10.1162/jocn_a_01990

later translated to humans (de Wit, Corlett, Aitken, Dickinson, & Fletcher, 2009; de Wit, Niry, Wariyar, Aitken, & Dickinson, 2007; Valentin, Dickinson, & O'Doherty, 2007). In this task, participants first learn to make a response to obtain a reward. Subsequently, the value of the outcome associated with that response is devalued, and the ability to flexibly adapt responding to this change in outcome value is measured during an extinction test. Sensitivity to outcome devaluation suggests that behavior is based on knowledge and evaluation of their consequences, and therefore under goal-directed control. If implementation intentions lead to "instant habits," then we would predict reduced sensitivity to outcome devaluation, reflecting a shift from goal-directed toward more rigid, habitual control (de Wit et al., 2018; Balleine & O'Doherty, 2010).

We have previously tested this hypothesis (van Timmeren & de Wit, 2022), using a computerized symmetrical outcome-revaluation task (SORT; Watson, Gladwin, Verhoeven, & de Wit, 2022). Participants learn to make a response (go) to certain ice cream vans to collect valuable ice creams (and points) or to withhold a response (no-go) to other ice cream vans delivering nonvaluable ice creams (and a reduction of points). To investigate the effect of if–then planning, we instructed them to use verbal implementation intentions for half of the stimuli and use goal intentions for the other half. In the subsequent test phase, some outcome values changed (i.e., outcome revaluation). Whereas participants should continue to respond according to the learned S–R mappings on value-congruent trials (i.e., still-valuable and still-not-valuable), they should flexibly adjust their behavior on value-incongruent trials (i.e., devalued and upvalued). The results of this previous study suggest that the use of implementation (compared with goal) intentions facilitates instrumental learning, but also impairs performance when some of the signaled outcome values change during the test phase (van Timmeren & de Wit, 2022). This detrimental effect of if–then planning was observed across value-congruent and incongruent trials, suggesting that it was not mediated by strengthened S–R associations (as this would have impacted the value-incongruent trials specifically). Instead, this result may have been driven by reduced goal-directed control. Investigating the neural processes underlying implementation intentions may offer us a window on the underlying (goal-directed vs. habitual) processes.

To this end, in the present study, we used fMRI to investigate the neural correlates of if–then planning of instrumental responses on the SORT. We capitalized on current insights regarding the neural basis of goal-directed and habitual control to investigate the notion that if–then planning gives rise to "instant habits." Decades of animal research have provided detailed insights into the neurobiology of goal-directed and habitual actions, demonstrating that they are causally supported by anatomically distinct but interacting corticostriatal systems (Balleine, 2019; Balleine & O'Doherty, 2010; Yin, Knowlton, & Balleine, 2004). These findings are mirrored by (correlational) neuroimaging evidence in humans, albeit less consistently. Specifically, previous fMRI studies have found that goal-directed control is supported by the ventromedial prefrontal cortex (vMPFC) and caudate whereas outcome-insensitive habitual actions depend on the premotor cortex and posterior putamen/dorsal striatum (Watson, van Wingen, & de Wit, 2018; Delorme et al., 2016; Morris, Quail, Griffiths, Green, & Balleine, 2015; de Wit et al., 2012; Tricomi, Balleine, & O'Doherty, 2009; Valentin et al., 2007).

The present study is the first fMRI investigation with the SORT, and we will therefore start with specifying our predictions regarding the general pattern of neural activity independent of intentions. First, we expected that over the course of training (i.e., habit acquisition) activity would increase in regions associated with habitual control whereas the involvement of regions implicated in goal-directed control would decrease (Zwosta, Ruge, Goschke, & Wolfensteller, 2018; Liljeholm, Dunne, & O'Doherty, 2015; Tricomi et al., 2009). Second, we expected neural activity during training in these regions to be predictive of revaluation insensitivity in the test phase (Watson et al., 2018; Zwosta et al., 2018; Liljeholm et al., 2015; de Wit et al., 2009). Third, in line with previous work (Watson et al., 2018; Valentin et al., 2007), we hypothesized that, in the test phase, we would find higher activity in areas implicated in goal-directed action, cognitive control, and response conflict when participants flexibly updated their responses and equal (if anything reduced) activity in habit-related regions. Finally, we expected that "slips of action" would be associated with higher activity in habit regions and reduced activity in goal-directed regions (Watson et al., 2018).

Our central aim was to investigate the neural basis of implementation intentions and their effect on behavioral flexibility. To this end, we measured neural activity related to the effect of implementation intentions on acquisition and flexible adjustment of instrumental actions on the SORT. We hypothesized that the use of implementation intentions (compared with goal intentions) during training would lead to increased habit acquisition as reflected by higher accuracy, increased automaticity (measured with the Self-Reported Behavioral Automaticity Index; Gardner, Abraham, Lally, & de Bruijn, 2012), and increased brain activity in habit regions and equal—or if anything reduced—activity in goal-directed regions. Moreover, we expected if–then planning to lead to increased reliance on previously formed S–R associations in the subsequent test phase as indicated by inflexible, habitual responding on value-incongruent compared with value-congruent trials, and higher activity of habit regions during the test phase. Finally, we expected that overcoming mentally rehearsed S–R associations (as part of an if–then plan) would require more goal-directed control and correspondingly engage related neural regions.

METHODS

All operationalizations, exclusion criteria, and main hypotheses and analyses were preregistered on Open Science Framework (https://osf.io/yrpxa).

Participants

Participants were recruited through the participant portal of the University of Amsterdam Web site, flyers, and word of mouth. We used the following inclusion criteria: age 16–35 years, not having previously participated in a previous study using this same task, and any contraindications for MRI. Data collection took place between July and November 2020. Note that this is during the first year of the COVID-19 outbreak; however, no strict lockdowns were implemented during this period in The Netherlands. The study was approved by the Psychology ethics committee of the University of Amsterdam and performed in accordance with those guidelines. All participants gave informed consent and received either course credit or financial compensation (15 €/hr) for their time (total ~2 hr). An additional €20 voucher was given to the participant with the highest score to motivate participants to perform well on the task.

Forty-seven participants were enrolled, conforming to our preregistered sampling plan. Our sample size was based on a previous pilot study, which found a significant effect of implementation intentions in 35 participants using the same task and manipulation. Moreover, a power analysis with G*Power (Version 3.1.9.3) showed that our target sample size of $n = 40$ should be sufficient to detect a small behavioral effect ($f = 0.12$) with an α level of .05 and power of .8. Six participants were excluded from all analyses. One participant quit half-way through participation, and five participants were excluded based on performance exclusion criteria (see Results for details). The remaining 41 participants (22 women, 19 men) had a mean age of 23.2 ($SD = 4.1$) years. All participants had normal or corrected-to-normal vision, and all were right-handed except one who was ambidextrous. All participants were free of neurological or psychiatric disorders and completed or were enrolled in higher professional education at the time of participation, the vast majority being university students. Two participants were native Germans who spoke Dutch fluently; all others were native Dutch speakers.

Stimuli and Materials

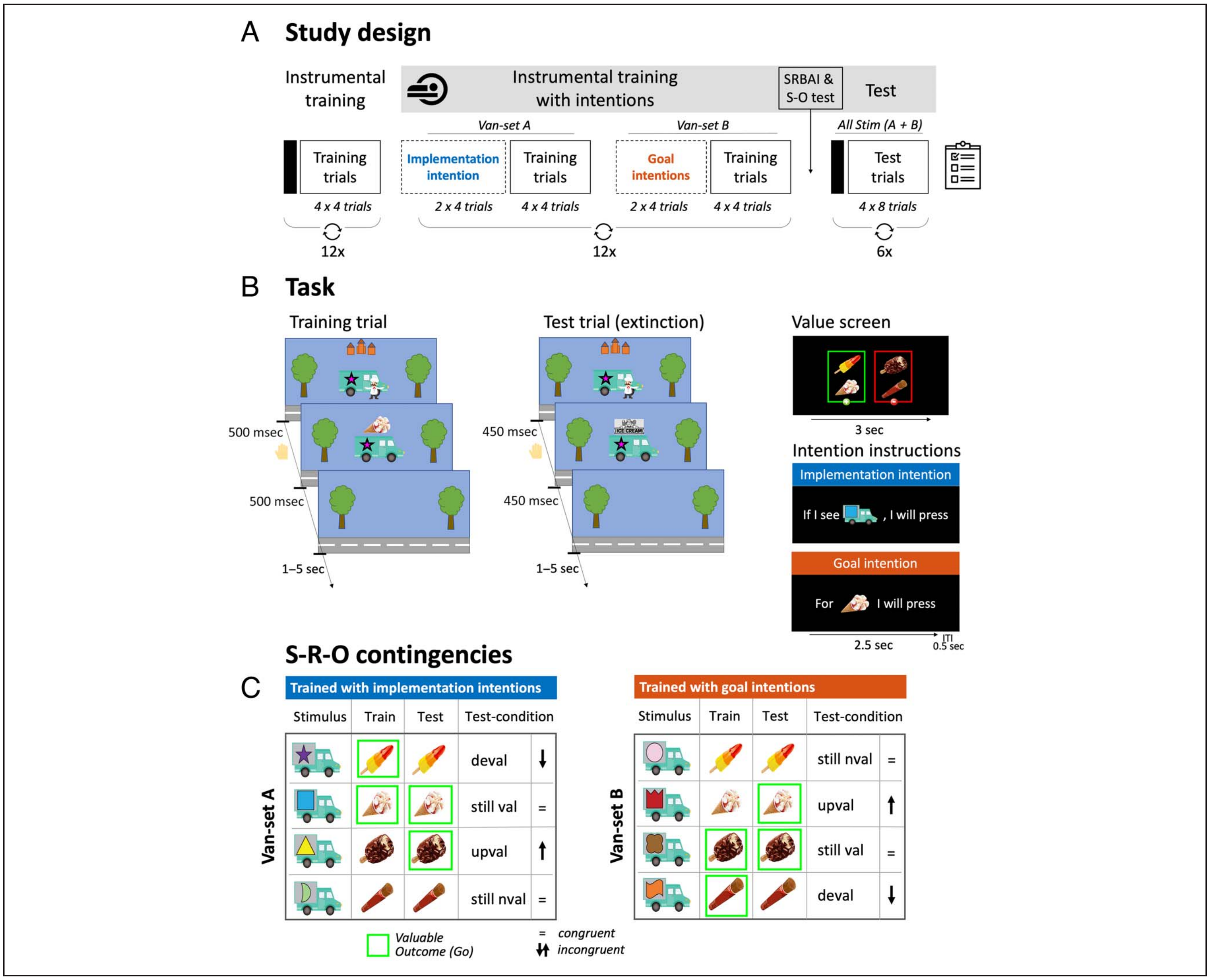
Procedure

Participants performed a computerized instrumental learning task called the SORT (Figure 1; Watson, Gladwin, et al., 2022), programmed in Presentation (Version 18.1). Participants played a hungry skateboarder with the objective to collect ice creams to earn points and satisfy their hunger by pressing a response button. They were informed that the best performing participant at the end of the study would receive a €20 voucher. Four pictures of ice creams were used: a Cornetto, a Magnum, a Rocket ice lolly, and a soft serve ice cream. The task consisted of three phases. First, participants conducted an instrumental training phase without strategic planning outside the scanner, after which they were moved to the MRI scanner and performed an instrumental training phase with strategic planning followed by a test phase (see Figure 1). The symmetrical nature of the task stems from the inclusion of both valuable and nonvaluable outcomes, which allows comparisons in the test phase (when outcome values change) between the value-congruent and value-incongruent conditions to be made with the same response type (see Watson, Gladwin, et al., 2022, for a more elaborate discussion on the advantages of this task). The total experiment took ~2 hr, of which 1 hr was spent in the scanner.

The task used here is almost identical to a previous study in which we tested the same hypothesis behaviorally (van Timmeren & de Wit, 2022), apart from the following changes. To minimize head movements, we used a static version of the task here instead of having ice cream trucks moving across the screen. We added one block of practice with strategic planning before being moved to the scanner, in order for participants to once read the intentions out loud and be able to ask questions. Moreover, we adapted the task to promote stimulus–outcome (S-O) learning across intention conditions, to rule out that any effect of implementation intentions on behavioral flexibility would be mediated by reduced contingency knowledge, as was the case in the original behavioral study (van Timmeren & de Wit, 2022). To this end, we changed the way in which the blocks were composed in the first part of training (i.e., without intentions): Instead of alternating between two sets of four ice cream vans, each block now contains four (out of eight) pseudorandomly selected stimuli (see Instrumental Training section for details). More than with the block-sets, participants are now forced to pay attention to all outcomes in the value-screen and evaluate for which stimulus they should (not) make a response.

Instrumental Training

At the start of the task, participants were instructed that their goal was to collect valuable ice creams (which earn points and alleviate hunger) and avoid collecting nonvaluable ice creams (which lose points and cause stomach pain) by (not) responding to ice cream vans. There were four different ice creams, and before each block of instrumental training, participants were shown which two ice creams were valuable (in green) and which two ice creams were not valuable (in red; Figure 1A). The position of the valuable and nonvaluable ice creams (left/right) was counterbalanced across participants. Each ice cream was associated with two out of eight vans (Figure 1B): one van always predicting this ice cream as being valuable and the other as being nonvaluable. Each block contained only half of the vans: two associated with a valuable ice cream and two with a nonvaluable ice cream. Participants were told to find out



Stimulus	Train	Test	Test-condition	
			deval	↓
			still val	=
			upval	↑
			still nval	=

Stimulus	Train	Test	Test-condition	
			still nval	=
			upval	↑
			still val	=
			deval	↓

Figure 1. Overview of the study and experimental design. Participants were told they were playing a hungry skateboarder and their goal was to collect some ice creams and not others to earn points. (A) Participants first received instrumental training. Each block started with a value-screen (represented by the black rectangle), followed by a block of 16 training trials (see B). Each block contained four vans (pseudorandomly selected). Training then continued with participants additionally using implementation intentions (trained with Van-Set A) or goal intentions (trained with Van-Set B; see C), with intention instructions (see B) being presented before each instrumental learning block. Finally, participants completed six test blocks in which all eight vans (Van-Sets A and B) would appear intermixed and consequently the associated outcome-values of some vans changed compared with training (see C, comparing the "Train" vs. "Test" columns). (B) Train trial: When a van was presented, participants had to decide whether to make a response within 500 msec, after which the ice cream appeared (irrespective of a response) on top of the van for 500 msec. Test trial: identical to train blocks, but now (i) a banner appeared on top of the van instead of the ice cream to prevent feedback about the outcome (i.e., nominal extinction) and (ii) response time was reduced to 450 msec. Value screen: The outcome-value screen indicates which ice creams should (in green) and should not (in red) be collected. Intention instructions: Vans were trained with either implementation intentions, indicating for which ice cream *van* they should or should not make a response, or goal intentions, indicating for which *ice cream* they should (not) make a response. (C) An overview of stimulus-outcome contingencies (example set) and associated values across different phases of the task. The contingencies between each ice cream and van remained consistent throughout the whole task, but the *value* of each ice cream (and hence the associated response) was stable only during training. During the critical test phase, the associated outcome values changed (were incongruent) relative to the training value for half of the stimuli (indicated by arrows). This results in four conditions: still-valuable trials (valuable, congruent), upvalued trials (valuable, incongruent), still-not-valuable trials (nonvaluable, congruent) and devalued trials (nonvaluable, incongruent). For example, the first van always delivered a Rocket, which was valuable throughout training but no longer valuable during test (i.e., devalued). Shown here is an example of the contingencies in one of six test blocks; across the test phase, the correct response for each stimulus was equally often congruent and incongruent. Deval = Devalued; still val = still-valuable; upval = upvalued; still not = still-not-valuable trials.

by trial and error which ice cream truck delivered which ice cream, and that the S-O contingencies would remain the same throughout the whole task. Participants first practiced with different discriminative stimuli (scooters) and outcomes (pizzas) for two blocks to familiarize them with this procedure. As mentioned previously, the composition of the blocks (i.e., which four out of eight vans were presented during this block) was now pseudorandomized. The conditions described above allow for six unique combinations of four vans, which were presented twice each (order randomized) during this first part of training for 12 blocks. The contingencies between ice creams and vans and which of the ice creams was valuable/nonvaluable were randomized across participants.

Each stimulus was shown 4 times per block, constituting 16 trials. Trial order was randomized per eight trials, with each van being presented twice in the first and twice in the second half of a block. Each trial started with a jittered 1- to 5-sec intertrial interval. Participants were instructed that they should respond as quickly as possible and before the deliverer disappeared (after 500 msec). Irrespective of the response, the associated outcome was then presented for 500 msec. Thus, participants did not receive direct feedback about the accuracy of their response to balance the feedback provided for valuable and nonvaluable outcomes and to promote goal-directed (R-O) learning and S-O knowledge. Each block ended with a 3-sec feedback screen that displayed accuracy and late responses in that block and total number of points collected (Figure 1D).

Instrumental Training with Intentions

The next phase of training took part in the MRI scanner. Participants were told that instead of seeing which ice creams were valuable or nonvaluable, each block would now start with sentences that would help them perform well. These sentences came in two different forms (Figure 1D). *Goal intentions* indicated for each *ice cream* whether they should make a response (R-O), formulized as "If I see [picture of an ice cream], then I WILL press." *Implementation intentions* indicated for each *ice cream van* if they should make a response or not (S–R), formulized as "If I see [picture of an ice cream van] then I WILL (NOT) press." Each intention was presented for 2500 msec and twice per intention block (randomized order). Half of the stimuli were trained using goal and the other using implementation intentions. Each block of verbal intentions was directly followed by a block of instrumental training (identical to the previous phase) with the corresponding stimuli. Blocks now alternated between two sets of vans, one van-set being trained with implementation intentions (S1–S4, "Van-Set A") and one with goal intentions (S5–S8, "Van-Set B"). Whether the training started with an implementation or goal intention block was counterbalanced across participants. At the end of regular instrumental training and before being moved to the scanner, participants practiced each verbal intention without instrumental training for one block, followed by two blocks (one for each intention type) with instrumental training. During these first few practice blocks outside the scanner, participants were asked to read the intentions out loud. During the subsequent 24 blocks of training with intentions in the scanner, participants were instructed to subvocalize the intentions instead of reading them out loud to minimize head motion. Participants entered the scanner in a head-first supine position and were able to view the screen using a mirror attached to the head coil on which the task stimuli were presented. A button box allowed them to collect ice creams by responding using their right index finger.

At the end of training with intentions, participants completed a questionnaire on subjective automaticity (Self-Report Behavioral Automaticity Index [SRBAI]) and were tested on their S-O knowledge (details below; Figure 1E). We had planned to additionally obtain a (pre-intention) baseline measure of these questionnaires, but because of a programming error, they were presented *after* the practice blocks with intentions, making them unusable as a baseline measure.

Test Phase

Participants completed six test blocks. The test phase was similar to the first training phase (without intentions), but with some important differences. First, as intention blocks were no longer presented, value-screens were again shown at the start of each block, for the duration of 4 sec. Second, participants were told that the ice cream deliverers placed a banner on top of their van, blocking the view of the ice cream they delivered (i.e., nominal extinction). Because each van still kept on delivering the same ice cream as during training, they should base their choice on what they learned before. Third, the feedback screens presented at the end of each block no longer included information on the accuracy of their responses, but only the percentage of responses, nonresponses, and late responses. We did this to prevent outcome-based learning during the test phase. We explicitly instructed participants that each block contained an equal amount of valuable and nonvaluable outcomes so they knew they should aim for a 50%/50% distribution. Fourth, we shortened the response window to 450 msec to force rapid responding, which has been shown to boost the expression of habitual slips (Hardwick, Forrence, Krakauer, & Haith, 2019). However, because a lot of participants responded just after the 450-msec time limit, we decided to include responses up to 600 msec for both the behavioral and fMRI analysis to increase the number of included trials in the fMRI analyses. This change did not significantly impact the pattern of behavioral results, which was unsurprising as the test phase was conducted in extinction, meaning that no performance feedback was provided during this period. Finally and crucially, participants were informed that the final phase would be more challenging because all eight ice cream vans would appear intermixed during each block. The crucial consequence of each block containing all eight stimuli is that half of the vans would now deliver an ice cream with a value incongruent with the value during training. Some ice cream vans for which they had been trained to always make a go response during training, now delivered a (devalued) ice cream that should not be collected. Vice versa, other ice creams vans had carried nonvaluable outcomes during training, but their signaled outcome was upvalued and therefore required a go response. On other (value-congruent) test trials, the signaled outcome remained the same (i.e., still-valuable and still-not-valuable trials).

Consider for example the Rocket ice cream in Figure 1C. In this example, during training, this ice cream is always delivered by the van with a purple star and the van with the pink circle. During training blocks with the van with the purple star, the Rocket is valuable and therefore requires a go response. In contrast, during training blocks with the van with the pink circle, the Rocket is not valuable, and participants should refrain from pressing the space bar (i.e., no-go response). Subsequently, during the test block all (ice cream van), stimuli are presented, and in the example illustrated in Figure 1C, the Rocket is instructed to be currently not valuable. This means that the van with the purple star signals a devalued outcome (i.e., this is value-incongruent with training and requires a different response), and the van with the pink circle signals a still-valuable outcome (i.e., value congruent; the learned response remains correct).

SRBAI

The SRBAI (Gardner et al., 2012) is a 4-item scale that captures self-reported habitual behavior patterns that we adapted for to assess automaticity for (not) responding to the ice cream vans. Participants were presented with each ice cream van and asked to indicate the associated response (press or not press) and the degree to which (not) making a response was something they did: "automatically," "without having to consciously remember," "without thinking," and "before I realize I am doing it." Each item was scored on a scale ranging from 1 (*strongly disagree*) to 100 (*strongly agree*). The SRBAI scale was previously shown to have good reliability and validity (Gardner et al., 2012). Before the four SRBAI items appeared, participants were asked to indicate which response was associated with that stimulus ("making a response" / "not making a response") to test S–R knowledge. Cronbach's alpha was calculated separately for each of the four conditions (2 intentions × 2 values), using the eight test items (four SRBAI questions for the two stimuli per condition). The results indicate high internal reliability, with alpha ranging from .91 to .95. The final score was calculated separately for each intention by taking the mean across the four items (range: 1–100), with higher scores reflecting more automatic behavior.

Test of Stimulus–Outcome Knowledge

Participants were asked about their knowledge of the S-O contingencies by asking them for each ice cream vans which ice cream it delivered. After selecting one of the four ice creams, participants were asked to indicate how confident they were about their decision (0–100). Composite scores, reflecting S-O knowledge, were calculated for each intention and separately for go- and no-go-trained stimuli by multiplying percentage of correct S-O contingencies (0%/50%/100%) with percentage mean confidence.

Preregistered Behavioral Data Analysis

Behavioral data analyses were performed using IBM SPSS Statistics 25 for Mac for frequentist statistics and JASP Version 0.16.3 (JASP Team, 2018) for Bayesian statistics. For data analysis purposes, the training data were collapsed across blocks of three, referred to as block-sets. Accuracy is reflected by the percentage of trials on which a correct response was made, calculated by the number of correct responses divided by the total number of trials. In line with the fMRI analyses, trials on which a late response was made were not included in the analyses (of both accuracy and RTs). To assess that learning took place over the first part of the training without intentions, accuracy was analyzed using a 2 × 4 repeated-measures ANOVA with within-subject factors Value (valuable or nonvaluable) and Block-set (1–4). The second part of training was analyzed using a 2 × 2 × 4 repeated-measures ANOVA, with Intention Type (implementation or goal intention) as an additional factor. RTs for correct responses (and thus only for valuable go trials) were analyzed with similar ANOVAs.

For the test phase, data were analyzed using a 2 × 2 × 2 repeated-measures ANOVA with three factors: Intention Type (implementation or goal intention), Test Value (valuable or nonvaluable during test), and Congruency (congruent or incongruent with value during training). Thus, for each intention type there are four conditions: still-valuable trials (valuable, congruent), upvalued trials (valuable, incongruent), still-not-valuable trials (nonvaluable, congruent), and devalued trials (nonvaluable, incongruent). Again, RTs (including all responses up to 600 msec) were analyzed using similar ANOVAs but now also analyzing responses on no-go trials (i.e., responses on still-not-valuable and devalued trials). Note that eight participants were excluded from the no-go analyses because they performed perfectly on still-not-valuable trials and thus did not make any response.

Subjective automaticity (SRBAI scores) for responding to stimuli trained with implementation and goal intentions at the end of training was compared using a paired t test. Finally, the relationship between automaticity and the "revaluation insensitivity" index was tested for both intention types separately using correlational analyses. A revaluation insensitivity index was calculated for each intention type by taking the difference between accuracy for congruent and incongruent test trials separately for go (still-valuable minus devalued) and no-go-trained stimuli (still-not-valuable minus upvalued), with higher revaluation insensitivity scores indicating more habitual performance. Kendall's tau was used as the four revaluation indices, and SRBAI scores were not normally distributed. In the case of violations of sphericity, we report Greenhouse–Geisser corrected degrees of freedom and p values. In addition to 95% confidence intervals, partial eta squared (η_p^2) for the ANOVAs and Cohen's d for paired t tests are reported as estimates of effect sizes.

We additionally conducted corresponding Bayesian analyses. For null results ($p > .05$), as preregistered, we report the Bayes Factor$_{01}$ (BF$_{01}$), which quantifies the relative evidence in favor of the null hypothesis (H0) over the alternative hypothesis (H1). For ANOVAs, we report the BF$_{\text{excl}}$, which quantifies the extent to which the data support inclusion of the factor of interest in the model (i.e., the change from before posterior inclusion odds, across matched models). Finally, although we interpret significant findings on the basis of $p < .05$, we also report BFs for comprehensiveness and transparency (i.e., BF$_{10}$ or BF$_{\text{incl}}$ for ANOVAs, which quantify evidence in favor of the alternative hypothesis over H0 and is identical to 1/BF$_{01/\text{excl}}$). BFs were interpreted according to Table 1 in Wetzels and colleagues (2011), with BFs between one and three reflecting anecdotal support, BFs larger than three reflecting substantial support, and BFs larger than 10 reflecting strong support. In all Bayesian analyses, JASP's default priors (cauchy = 0.707 for t tests and r = 0.5 for fixed and r = 1 for random effects for ANOVAs) were used.

MRI Data Acquisition

All MRIs were performed on a 3-Tesla, full-body Achieva dStream MRI-scanner (Philips Medical Systems) equipped with a 32-channel head coil. After entering the scanner, a low-resolution survey scan was made to determine the location of the field of view.

fMRI scans were acquired at a ~30° angle from the anterior–posterior commissure line to maximize signal sensitivity in orbital regions (Deichmann, Gottfried, Hutton, & Turner, 2003) using a T2*-weighted single-shot gradient echo imaging sequence with the following parameters: repetition time = 2000 msec; echo time = 28 msec; flip angle = 76.1°; voxel size = 3 mm^3 with 0.3-mm slice gap; matrix size = 80 × 78; number of slices = 36; field of view = 240 × 118.5 × 240 mm. The training with intentions was split in two runs of 598 scans each, whereas 415 scans were acquired for the test phase. The first six volumes of each run were discarded to allow T1 saturation to reach equilibrium.

A high-resolution T1-weighted structural image was acquired before the final run (while participants completed the post-training SRBAI and SO-test) using an MPRAGE sequence with the following parameters: voxel size = 1 mm^3; field of view = 240 × 220 × 188 mm; repetition time = 8.2 msec; echo time = 3.7 msec, 220 slices, flip angle = 8°.

fMRI Data Analysis

Image Preprocessing

MRI data were first converted to Brain Imaging Data Structure format using in-house scripts. An initial check of data quality was done by visually inspecting the image-quality metrics derived from MRIQC v0.15.0 (Esteban et al., 2017). Data were preprocessed using fMRIPrep v20.1.1 (Esteban et al., 2019; RRID:SCR_016216), which is based on Nipype 1.5.0 (Gorgolewski et al., 2011; RRID: SCR_002502), with the default processing steps. These included brain extraction, segmentation, and surface reconstruction of the structural T1 image; spatial normalization of both the structural and functional data to MNI space; and head motion estimation, coregistration, susceptibility distortion correction, and resampling to 2 mm^3 of the functional data. No slice-timing correction was performed. A comprehensive description of the preprocessing pipeline is available here: https://osf.io/72bsh.

fMRI Statistical Analyses

The preprocessed functional data were further analyzed using Statistical Parametric Mapping software (SPM12,

Table 1. Imaging Results of the Training Phase (Exploratory)

Contrast	*Region*	*MNI Coordinates (x, y, z)*			*Cluster Size (Voxels)*	*z Score at Peak Level*	*Correction*
Increase over training blocks (go)	Caudate nucleus head	22	6	30	443	4.37	Cluster
	Amygdalo-hippocampal junction	−10	−4	−14	348	5.17	Peak
	Angular gyrus	20	−52	38	214	4.92	Peak
	Posterior putamen	26	−20	4	34	3.96	SVC Tricomi
Decrease over training blocks (go)	Anterior caudate L	−24	10	2	912	6.53	Cluster
	Anterior caudate R	24	10	−4	537	6.39	Cluster
	Primary motor/SMA	8	−24	60	860	5.44	Cluster
	Hippocampus/putamen	43	14	−8	657	4.83	Cluster
	Temporal cortex L	−46	−46	−4	591	5.69	Cluster
Goal > implementation intentions block-set 1 (Go)	Anterior caudate	13	18	−4	40	3.69	SVC striatum

SVC = small volume correction; L = left; R = right.

Wellcome Trust Centre for Neuroimaging). The data were spatially smoothed using a Gaussian kernel with a FWHM of 8 mm and all functional data was high pass filtered (with a 128-sec cutoff) to remove slow signal drifts.

First-Level Analysis

For the first-level analysis of the fMRI data, a general linear model was constructed for each participant, concatenated over all three runs from the training and test phase. For the data on training with intentions, trial onsets of valuable stimuli and nonvaluable stimuli for implementation and goal intentions were modeled using stick functions, making four conditions. To look at the effect of time on training, these were modeled as separate regressors per three blocks, making four training block-sets. Only correct trials (i.e., where an accurate (non)response was made) were included. Blocks of verbal rehearsal of implementation and goal intentions were additionally modeled as blocks of 28 sec (total duration of eight 3.5-sec trials). For the test phase, stick functions modeled the trial onsets of still-valuable and still-not-valuable ("value-congruent"; the outcome value is congruent with training phase) and devalued and upvalued ("value-incongruent"; the outcome value is not congruent with training phase) stimuli that were trained with implementation or goal intentions separately, making eight regressors. To investigate BOLD activity during habitual (c)omission errors (habitual "slips" in case of incongruent trials), separate regressors were included for incorrect trials for all conditions. The following regressors of no interest were included separately for each run: one regressor for errors (only for training, as test-errors/"slips" were modeled as regressors of interest) and late trials, keypresses, feedback-displays, value-screens (only for test phase), and six realignment parameters capturing rotation and translation to correct for residual participant motion. Three session constants were included in the model. All onsets were then convolved with the canonical hemodynamic response function, and an autoregressive AR(1) model was used to correct for serial correlations. The general linear model was regressed against the fMRI data to generate parameter estimates for each participant.

Regressor-specific first-level contrast images were created for the training- and test-regressors modeling the different conditions of interest to construct the planned second-level full factorial models. These contrasts of parameter estimates were then entered into between-subjects ANOVAs to generate group-level random-effects statistics. To test for a difference in learning between intention types, contrasts of parameter estimates of the instrumental training phase were entered into a 2 × 4 × 2 (Value × Block-set × Intention) factorial ANOVA. Following estimation of the second-level model, t tests were specified by adding linear weights to each instrumental training block-set, modeling increases over training as $[-1.5\ -0.5\ 0.5\ 1.5]$ and decreases as $[1.5\ 0.5\ -0.5\ -1.5]$.

In addition, first-level contrast images were created. To assess the effect of planning during training, contrasts were created comparing training with implementation versus goal intentions (across all blocks, separately for go and no-go trials). To examine markers of goal-directed control during test, we compared correct congruent trials with correct incongruent trials (i.e., [still-valuable go > upvalued go] and [still-not-valuable no-go > devalued no-go]). We also investigated situations where participants fail to adapt to the new outcome value and continue to respond according to the learned S–R association by comparing incorrect incongruent trials (i.e., "slips of action") with correct incongruent trials. Again, separate contrasts were created for test-go- and test-no-go trials (i.e., [devalued go > upvalued go] and [upvalued no-go > devalued no-go]). Finally, we also created a similar contrast comparing incorrect incongruent trials (slips) with correct congruent trials (i.e., [devalued go > still-valuable go] and [upvalued no-go > still-not-valuable no-go]). More information about the rationale behind these contrasts is provided in the Results section. To assess the effect of planning strategy on test performance, the same test-phase contrasts were constructed but looking for an interaction with intention type (e.g., [still-valuable go > upvalued go × implementation > goal intention]). Parameter estimates generated from these first-level analyses were entered into a random-effects group analysis, and linear contrasts were used to identify significant effects at the group level.

Higher level whole-brain statistical maps were corrected for FWE at the cluster-level ($p_{\text{FWE-cluster}} < .05$) with a voxel cluster-defining threshold of $p = .001$ uncorrected. When activations did not reach statistical significance at the cluster level, we also checked the peak-voxel level with a threshold of $p < .05$ corrected ($p_{\text{FWE-peak}} < .05$). In such cases, we clearly indicate this in the text, and we report the peak-voxel level results so as to be as comprehensive as possible in our reporting. Finally, in an exploratory analysis, we further aimed to test for effects in specific regions of the striatum given prior published findings on the role of these structures in goal-directed and habitual responding (Watson et al., 2018; de Wit et al., 2012; Tricomi et al., 2009; Tanaka, Balleine, & O'Doherty, 2008; Valentin et al., 2007). In particular, we defined an anatomical ROI to examine effects in the caudate nucleus, a region previously implicated in goal-directed processes, as well defining a functional ROI based on the results from Tricomi et al. (2009) that implicated the posterior putamen in habit-related processing.

In addition, we identified several ROIs in our preregistration: for habitual control, goal-directed control, response conflict, and implementation intentions. Three separate masks were created based on these ROIs to apply small volume correction (SVC). Apart from a striatal ROI (encompassing the bilateral caudate, putamen, and NAcc from the AAL atlas (Tzourio-Mazoyer et al., 2002); however, applying SVC with the three preregistered ROIs did not alter the pattern of results. This may be because of the large number of voxels included in the ROIs (especially the goal-directed mask) thereby reducing the sensitivity of the SVC. Therefore, we have opted to stick to reporting the whole-brain results for the confirmatory analyses. Whole-brain *t*-maps (without thresholding) of the main fMRI contrasts are available at https://neurovault.org/collections/13191/.

RESULTS

All analyses reported in this section were preregistered at the start of this study, unless indicated otherwise in the text. We generally followed the preregistered analysis plan, but in some cases, the results prompted us to further explore the data. We should also point out that we preregistered these hypotheses before finishing data analysis of our related behavioral study (van Timmeren & de Wit, 2022). Hence, we preregistered the same behavioral hypotheses for this study, although the original behavioral study only partially supported our initial predictions—a point we will come back to in the discussion. We therefore incidentally deviate from the preregistration to keep our analyses in line with analyses and findings from the behavioral study, which is always clearly indicated.

The total final sample used for the analyses consisted of 41 participants, after excluding the following participants. On the basis of the preregistered exclusion criteria, no participants were excluded on the training criterion (< 80% accuracy in the last block-set of training), whereas three were excluded because they made < 25% responses on upvalued trials trained with goal intentions in the test phase. The goal of this criterion was to ensure that participants understood the test-phase instructions and updated their performance accordingly, while not excluding participants based on the manipulation of interest (i.e. implementation intentions). We additionally excluded two participants (post hoc) based on a very low overall response rate during the test phase. Although these participants made (just) > 25% upvalued responses, we deviated from the preregistration because they were outliers on the overall response rate and responded on less than one out of three trials during the test, despite receiving explicit instruction to aim for a response rate of ~50% and receiving feedback about that at the end of each block. Hence, they did not follow the test-phase instructions and their performance is not reliable. Note that this criterion is independent of actual task performance (accuracy) and that the in-/exclusion of these two participants does not change the general pattern of behavioral nor fMRI results.

Behavioral Results

Training Phase without Intentions

As expected, participants learned to make correct responses over the first part of training (Figure 2A), as revealed by a significant main effect of Block-set on accuracy, $F(2.46, 98.20) = 16.74, p < .001\ \eta_p^2 = .30$, $BF_{incl} = 2.81 \times 10^5$, and a marginally significant effect of Block-set on RT, $F(2.45, 98.07) = 2.75, p = .058, \eta_p^2 = .06$, $BF_{incl} = 0.81$. There was no significant difference in learning to make go versus no-go responses (main effect of Value: $F(1, 40) = 2.00, p = .17, \eta_p^2 = .05$, $BF_{excl} = 1.60$; Block × Value interaction: $F(1.70, 68.16) = .25, p =. 57, \eta_p^2 = .01$, $BF_{excl} = 22.15$).

Instrumental Training with Goal versus Implementation Intentions

Following the first 12 blocks of instrumental training without planning, intentions were introduced during a practice block (still outside the scanner). Although we did not preregister to analyze those data, for completeness and in line with our previous behavioral study with this paradigm investigating the same question (van Timmeren & de Wit, 2022), we conducted a paired *t* test comparing the final block of training without intentions to the practice block. This analysis revealed that participants benefitted from if–then planning on the valuable go trials, as reflected by higher accuracy ($M = 96.1$, $SD = 12.4$) relative to the preceding (pre)training block-set (baseline: $M = 91.8$, $SD = 9.1$, $Z(40) = 2.57, p = .01, d = 0.59$, 95% CI [.81, .22], $BF_{10} = 1.34$), whereas RTs were not affected, $t(40) = -.01, p = .99, d = -0.001$. In contrast, the use of goal intentions negatively impacted both accuracy ($M = 87.6$, $SD = 14.7$, $Z(40) = -1.86, p = .065, d = -0.40$, 95% CI [−.01, −.69], $BF_{10} = 1.36$) and RTs, $t(40) = -2.03, p = .049, d = -0.32$, $BF_{10} = 1.08$, of go trials compared with (pre)training. For no-go trials, no significant effects of implementation, $Z(40) = 1.03, p = .31$, $BF_{01} = 5.12$, or goal intention, $Z(40) = .10, p = .93$, $BF_{01} = 5.68$, were seen.

Subsequently, when instrumental training was resumed during the scanning session, the 2 × 2 × 2 repeated-measures ANOVA indicated that the advantage of if–then planning was initially still apparent on valuable go trials (Figure 2A). In addition to a strong main effect of Value, driven by participants performing better overall on valuable compared with nonvaluable trials, $F(1, 84.47) = 10.93, p = .002, \eta_p^2 = .22$, $BF_{incl} = 18.08$, we found the expected preregistered three-way interaction between Intention, Value, and Block-set, $F(3, 103.14) = 6.45, p < .001, \eta_p^2 = .14$, $BF_{incl} = 857.7$. Separate analyses of valuable and nonvaluable trials revealed a significant Intention × Block interaction for valuable, $F(3, 81.78) = 6.21, p = .003, \eta_p^2 = .13$, $BF_{incl} = 74.01$, but not for nonvaluable trials, $F(3, 120) = 1.88, p = .14, \eta_p^2 = .05$, $BF_{excl} = 2.63$. The significant effect on the valuable go trials was driven by higher accuracy with implementation compared with goal intentions during the first block-set, $Z(40) = 3.34, p < .001, d = 0.85$, 95% CI [.64, .94], $BF_{10} = 22.76$. At the end of training (Block-Set 4), there was no longer a significant effect of Intention Type on accuracy, $Z(1, 40) = -.34, p = .80, \eta_p^2 = -1.43$, $BF_{01} = 5.87$.

The analysis of RTs (Figure 2A) revealed a main effect of Intention Type, $F(1, 40) = 12.08, p = .001, \eta_p^2 = .23$, $BF_{incl} = 11.12$, with faster responses during blocks trained with implementation intentions (median = 365 msec, SD =

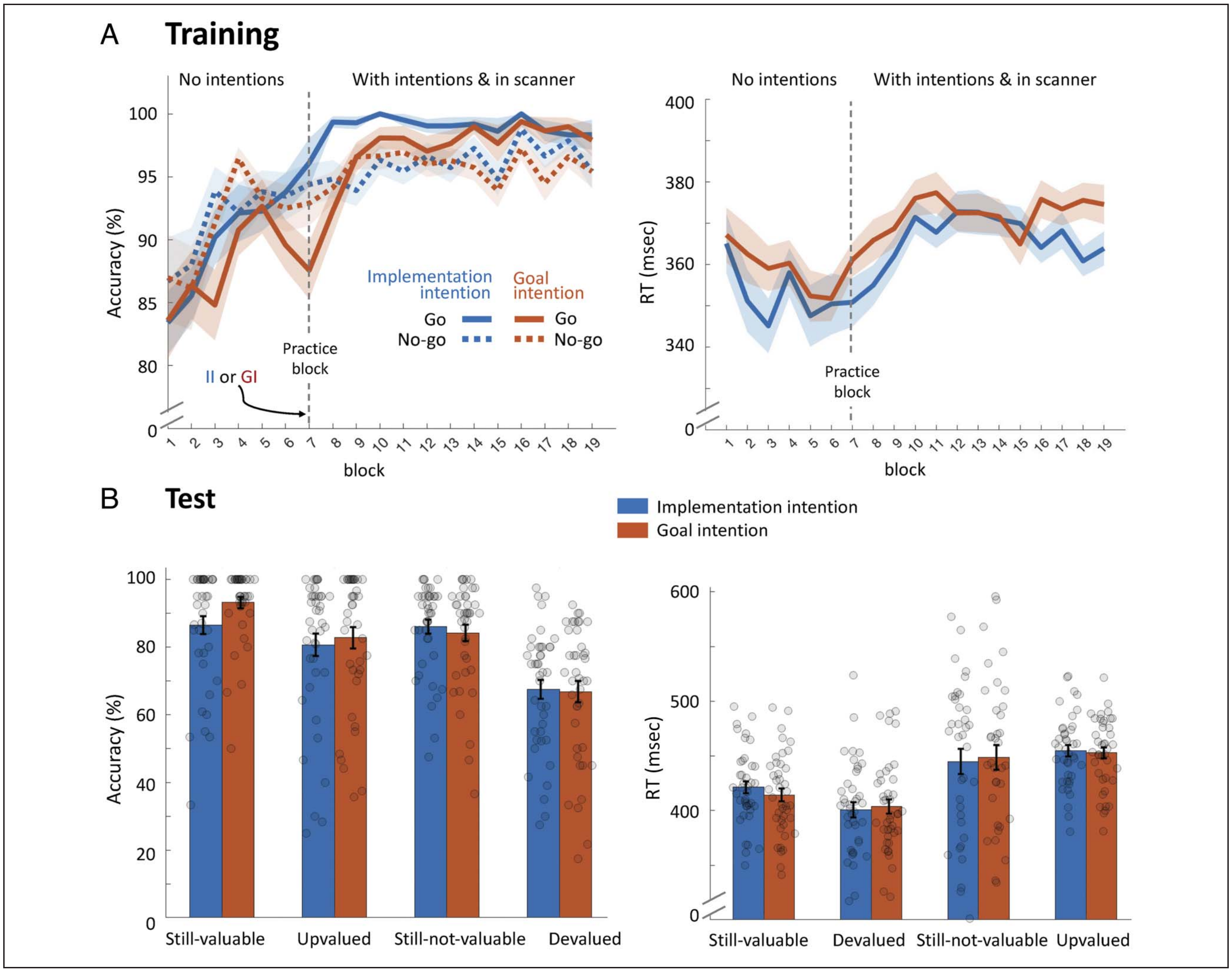


Figure 2. Behavioral results. (A) Over the course of training, participants learned to successfully respond for stimuli associated with valuable outcomes (Go) and to withhold making a response for stimuli associated with nonvaluable outcomes (no-go), as reflected by increasing accuracy rates. After six blocks of regular training, some stimuli continued to be trained using implementation intentions (blue) whereas others were trained with goal intentions (blue). Following one block of practice (black dotted line), participants were moved to the scanner and resumed training with intentions. Accuracy was significantly higher initially when using implementation intentions, but toward the end of training performance was almost perfect for both implementation and goal intentions. Across training with intentions, participants were faster during blocks trained with implementation versus goal intentions. (B) During the test phase, for some stimuli, the associated outcome changed in value (and thus response) compared with training (upvalued and devalued; see Figure 1C) and participants had to flexibly update their responses accordingly. For other stimuli, the associated value and response remained congruent with training (still-valuable and still-bot-valuable). Participants responded less accurately for incongruent compared with congruent trials, reflecting inflexibility as a consequence of learned S–R contingencies during training. However, training with implementation intentions did not lead to reduced flexibility. Similarly, there was no significant effect of training with implementation intentions on RT. (Shaded) error bars represent standard error of the mean. II = implementation intentions; GI = goal intentions.

17) compared with goal intentions (median = 374 msec, *SD* = 20), but no significant effect of Block-set, $F(2.4, 98.6) = 2.31$, $p = .09$, $\eta_p^2 = .05$, $BF_{excl} = 3.41$, nor an interaction ($p = .20$, $\eta_p^2 = .04$, $BF_{excl} = 3.67$).

Symmetrical Outcome-Revaluation Test

As expected, learned S–R associations had a clear impact on performance during the test phase (Figure 2B), as revealed by a main effect of Congruence, $F(1, 40) = 65.08$, $p < .001$, $\eta_p^2 = .62$, $BF_{incl} = 1.39 \times 10^7$. Because test value showed significant interactions with both Congruence, $F(1, 40) = 10.73$, $p = .002$, $\eta_p^2 = .21$, $BF_{incl} = 8.91$, and Intention Type, $F(1, 40) = 5.94$, $p = .02$, $\eta_p^2 = .13$, $BF_{incl} = 1.27$, separate follow-up comparisons were conducted for go (associated with still-valuable and upvalued outcomes) and no-go (associated with still-not-valuable and devalued outcomes) trials. Main effects of Congruence were seen for both the go, $F(1, 40) = 16.82$, $p < .001$, $\eta_p^2 = .30$,, $BF_{incl} = 76.40$, and no-go, $F(1, 40) = 56.46$, $p < .001$, $\eta_p^2 = .59$, $BF_{incl} = 2.31 \times 10^6$, stimuli. As can be seen in Figure 2B, the congruency effect was larger for no-go trials mainly because of participants struggling more on devalued trials, where they had to suppress responding to discriminative stimuli that previously signaled a valuable outcome. Importantly, we were interested in the effect of implementation intentions on test performance. First, an analysis of the go test trials suggested that overall performance was worse when trained with implementation compared with goal intentions, $F(1, 40) = 5.48, p = .02$, $\eta_p^2 = .12$, although Bayesian statistics showed that this evidence was inconclusive ($\text{BF}_{\text{incl}} = 1.46$). Importantly, in contrast to our preregistered hypothesis, there was no evidence for reduced flexibility as a consequence of if–then planning: The expected interaction of congruence with intention type failed to reach significance, $F(1, 40) = 1.52, p = .23$, $\eta_p^2 = .04$, $\text{BF}_{\text{excl}} = 1.86$. Given the direct relevance of the comparison between intentions for our research question, we followed these analyses up with separate (exploratory) paired *t* tests for still-valuable and upvalued trials to also report Bayesian evidence against a difference. Findings indicate that intentions only had a significant negative effect on (congruent) still-valuable, $Z(40) = -2.55, p = .01$, $d = -0.56$, $\text{BF}_{\text{incl}} = 3.68$, but not on (incongruent) upvalued trials, $t(40) = -.75, p = .46$, $\text{BF}_{\text{excl}} = 4.54$. Finally, for the no-go stimuli (still-not-valuable and devalued), no main, $F(1, 40) = .42, p = .52$, $\text{BF}_{\text{excl}} = 4.37$, nor interaction, $F(1, 40) = .06, p = .81$, $\text{BF}_{\text{excl}} = 4.25$, effects of intention type were observed.

We also analyzed RTs during the test phase. A Value × Congruence interaction, $F(1, 32) = 49.47, p < .001$, $\eta_p^2 = .61$, $\text{BF}_{\text{incl}} = 2.91 \times 10^5$, prompted separate analyses for trials trained with go responses (still-valuable and devalued) and for trials trained with no-go responses (still-not-valuable and upvalued). Interestingly, there was a main effect of congruence for go-trained stimuli, suggesting significantly faster RTs on devalued trials (*M* = 418 msec, *SE* = 8.8) relative to still-valuable (*M* = 443 msec, *SE* = 6.8; $F(1, 40) = 12.56, p = .001$, $\eta_p^2 = .24$, $\text{BF}_{\text{incl}} = 23.40$), in line with the idea that habitual slips of action are triggered fast and efficiently before one has the chance to suppress them. As late responses were excluded from this analysis (following the accuracy analysis), we ran an additional analysis including RTs for late responses to make sure that this effect was not driven by a higher number of (excluded) late responses on devalued trials. This analysis showed an even stronger main effect of congruence than the original analysis without late responses, $F(1, 40) = 14.84, p < .001$, $\eta_p^2 = .27$, $\text{BF}_{\text{incl}} = 36.88$. No other significant effects of RTs were found (all $p > .22$, $\text{BF}_{\text{excl}} > 1.74$).

Self-reported Automaticity and S-O Knowledge

Self-reported automaticity was at a high level overall (median = 80.4%, *SD* = 16.7), but did not differ between intentions, $t(40) = -.98, p = .34$, $\text{BF}_{01} = 3.80$, nor did subjective automaticity correlate with revaluation insensitivity for implementation ($r_\tau = -.09, p = .57$, $\text{BF}_{01} = 4.39$) or goal intentions ($r_\tau = .22, p = .17$, $\text{BF}_{01} = 2.03$).

Following van Timmeren and de Wit (2022), we also explored differences in S-O knowledge between intention types and their relationship with overall test accuracy. S-O knowledge was high (median = 89.8%, *SD* = 22.1) and, contrary to our previous study, no longer differed significantly between intention types, $F(1, 40) = 2.07, p = .16$, $\eta_p^2 = .05$, $\text{BF}_{01} = 2.6$; values, $F(1, 40) = 3{,}42, p = .07$, $\eta_p^2 = .08$, $\text{BF}_{01} = 2.4$; or their interaction, $F(1, 40) = .91, p = .35$, $\eta_p^2 = .02$, $\text{BF}_{01} = 5.88$, suggesting that the adaptation we made to the task (i.e., using a pseudorandom selection of stimuli instead of alternating between two block-sets in the first part of training, see Methods section) had the desired effect. S-O knowledge did correlate positively with test accuracy (across all four conditions) for both implementation intentions ($r_\tau = .30, p = .008$, 95% CI [.08, .52], $\text{BF}_{10} = 7.91$) and goal intentions ($r_\tau = .39, p < .001$, 95% CI [.21, .57], $\text{BF}_{10} = 99.22$).

Conclusions: Behavioral Results

We provide evidence for habit learning, as indicated by the general effect of previously learned S–R mappings on the ability to flexibly adapt responding when the cue signals a revalued outcome (i.e., incongruent). Importantly, although if–then planning seemed to increase efficiency relative to goal intentions, as reflected in superior acquisition, this was not at the expense of flexibility when outcome values changed in the test phase.

Neuroimaging Results

Instrumental Training: Across Intentions (Exploratory)

First, we were interested to explore general learning effects across intention types because this was the first time the SORT was used in the MRI scanner. These analyses showed that over the course of go training (i.e., on valuable trials), activity increased linearly in the head of the caudate nucleus extending into ACC (at $p < .05$ FWE rate corrected; $p_{\text{FWE-cluster}} < .05$). Activation in the left amygdalo-hippocampal junction and the angular gyrus did not reach our cluster-level correction threshold, but did survive voxel-level correction at $p < .05$ ($p_{\text{FWE-voxel}} < .05$; Table 1). In this same contrast, we also observed a cluster in the posterior putamen, which survived a small-volume correction for the posterior putamen ROI (i.e., $p_{\text{FWE}} < .05$ with SVC, defined as a 10-mm sphere at peak value of the cluster that showed a significant increase over training in the study of Tricomi et al. [2009]; $x = 33$, $y = -24$, $z = 0$). On the other hand, activity decreased over training in the bilateral anterior caudate (a more ventral part of the striatum), primary motor cortex (extending to mid-posterior cingulate), hippocampus extending into the putamen, and the left temporal cortex (all $p_{\text{FWE-cluster}} < 0.05$ corrected). In contrast, on no-go trials, there were no voxels that showed a significant linear change over training blocks.

Instrumental Training: Comparing Goal and Implementation Intentions

We then examined whether strategic planning affected instrumental training. The contrast comparing the average BOLD signal of trials trained with implementation intentions and goal intentions did not reveal any significant activations, neither on go nor no-go trials. We also tested for differences in learning between intentions over the course of training by adding linear weights to block-sets to compare increased activity over block-sets during implementation intentions with decreased activity during goal intentions, and vice versa. However, both tests of this interaction failed to show significant differences.

The finding that implementation intentions showed the most pronounced effect behaviorally early in training prompted us to conduct an exploratory analysis of only the first training block-set. This analysis revealed significantly decreased activation in the anterior caudate ($p_{\text{FWE}} < .05$ with SVC, $z = 3.69$) on trials trained with implementation intentions compared with goal intentions (Figure 3A and Table 1). For visual purposes, the extracted average BOLD signal from the anterior caudate cluster is shown separately for each block-set and intention in Figure 3B. As can be seen here, activity was indeed lower on implementation intention trials during the first block-set only and subsequently decreased for both intentions. A whole-brain analysis also showed decreased activity for implementation relative to goal intentions at an uncorrected threshold ($p < .001$) in the right lateral orbitofrontal cortex (OFC; $p_{\text{FWE-cluster}} = .061, z = 4.25; x = 26, y = 50, z = 14$) and the left insula ($p_{\text{FWE-cluster}} = .28, z = 3.76; x = -42, y = 20, z = 2$). However, because these results did not survive FWE-correction, we refrain from interpreting them further. To rule out that these findings were driven by RTs, which were significantly shorter for implementation compared with goal intentions, we performed an additional analysis controlling for trial-by-trial RT by including a parametric regressor (one for each of the two training runs) with RTs for each trial. This had no significant impact on the results, and we could qualitatively replicate all reported findings.

Neural Predictors of Test Performance

To determine whether brain activity during instrumental training with implementation intentions was predictive of test performance, we tested whether the average BOLD signal during training covaried with the revaluation insensitivity score. This preregistered test did not reveal significant neural predictors of test performance. For completeness, we also exploratively ran this analysis separately for goal intentions and across intentions, but this similarly did not reveal any significant results.

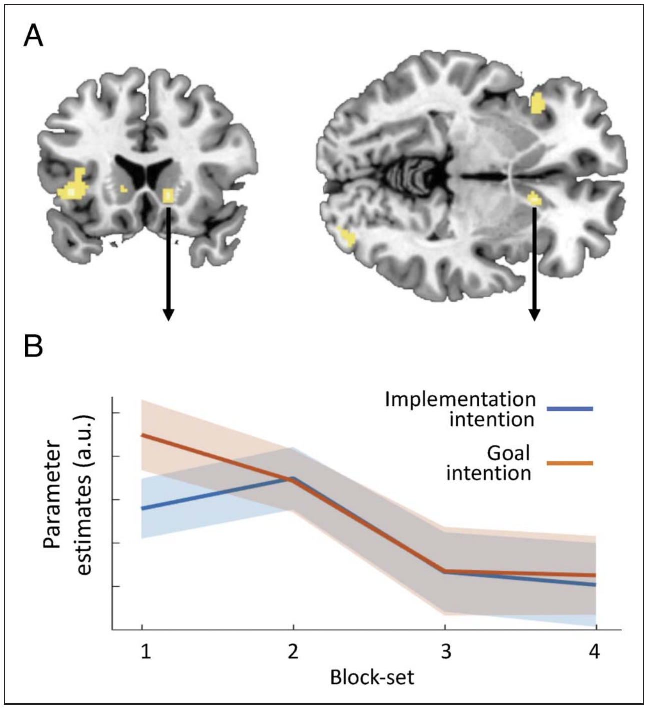


Figure 3. Lower activity in the right anterior caudate early in training for implementation compared with goal intentions. (A) Voxels that showed significantly lower activation during the first block-set of training with implementation compared with goal intentions on go-trials (at $p_{\text{FWE}} < .05$, small volume-corrected). The activity patterns shown are thresholded at $p < .001$ uncorrected. (B) Parameter estimates extracted from this anterior caudate cluster (peak at $x = 13, y = 18, z = -4$) over block-sets. Error bars represent 95% confidence intervals. a.u. = arbitrary units.

Symmetrical Outcome-Revaluation Test: Markers of Goal-directed versus Habitual Performance

In the test phase, changes in outcome value create conflict between goal-directed control and learned S–R associations. Specifically, to perform the correct response on incongruent trials (i.e., upvalued go and devalued no-go), participants have to exert goal-directed control and override the learned S–R mapping. Conversely, on congruent trials (still-valuable go and still-not-valuable no-go), participants can rely on the learned S–R associations. The advantage of the symmetrical outcome-revaluation test (compared with the original slips of action test) is that we can compare congruent and incongruent trials with each other unconfounded by test outcome value (and therefore required response: i.e., go or no-go). Therefore, to examine markers of goal-directed control, we firstly compared upvalued go with still-valuable go responses and found that this was associated with increased right insula activity ($p_{\text{FWE-cluster}} < .05$, $z = 4.16$; Table 2). No significant activations were seen in the contrast between devalued no-go and still-not-valuable no-go trials.

To identify regions where participants fail to adapt and continue to respond according to the learned S–R

Table 2. Imaging Results of the Test Phase

Contrast	*Region*	*MNI Coordinates (x, y, z) Max*			*Cluster Size (Voxels)*	*z Score (Peak)*	*Correction*
Upvalued go > still-valuable go	Insula R	38	24	−2	468	4.16	Cluster
Devalued slips > still-valuable go	Anterior insula L	−40	26	2	611	5.46	Cluster
	Anterior insula R	42	26	−10	621	4.49	Cluster
Still-valuable go > devalued slips	vMPFC	22	42	−4	388	4.64	Cluster
	Caudate	8	28	2			
	NAcc	4	20	4			
	Primary motor cortex	−26	12	60	252	5.10	Peak
	Paracentral lobule	−10	−30	66	336	4.56	Cluster
	Angular gyrus L	−30	−52	52	1653	5.51	Cluster
	IPL L						
	Angular gyrus R	38	−50	58	2510	5.33	Cluster
	IPL R						
	Occipital cortex	−36	−74	8	767	5.25	Cluster
Devalued slips > upvalued go	Anterior insula L	−36	26	−8	707	4.29	Cluster
	SMA	8	8	64	378	5.47	Cluster
	dACC	8	18	34	431	4.18	Cluster
	Inferior parietal lobule L	−56	−42	34	269	4.60	Peak
	Inferior parietal lobule R	56	−4	44	331	4.44	Cluster
	Supramarginal gyrus	−36	26	−8	707	4.29	Cluster
Upvalued go > devalued slips	ACC, caudate nucleus	−20	22	18	327	4.17	Cluster
	Premotor/PMC	−26	0	42	529	4.80	Cluster
	Lateral OFC	−32	62	0	317	4.19	Cluster
	Superior parietal love L	−28	−76	36	1482	4.12	Cluster
	Superior parietal love R	30	−62	38	4099	5.04	Cluster
	Occipital/visual cortex	−30	−96	16	1307	6.43	Cluster

L = left; R = right; NAcc = nucleus accumbens; IPL = inferior parietal lobule; (d)ACC = (dorsal) anterior cingulate cortex; PMC = primary motor cortex; OFC = orbitofrontal cortex.

association, we contrasted incorrect incongruent trials (devalued go and upvalued no-go) to correct incongruent trials (upvalued go and devalued no-go, respectively), as the latter arguably require most goal-directed control to override the learned S–R mapping. The contrast comparing devalued go responses (i.e., slips of action) with upvalued go responses is shown in Figure 4A, and revealed increased activity in a fronto-parietal network, including the left anterior insula extending to the inferior lateral prefrontal cortex, SMA, dorsal anterior cingulate cortex, bilateral inferior parietal lobule, and supramarginal gyrus (all $p_{\text{FWE-cluster}} < .05$; Table 2). Conversely, lower activity during slips of action compared with upvalued go responses was seen in the left anterior cingulate cortex extending into caudate nucleus, left lateral OFC, bilateral superior parietal lobe, and several occipital/primary visual areas (all $p_{\text{FWE-cluster}} < .05$, Table 2). In addition, activation in the premotor/primary motor cortex did not survive cluster-level correction but did reach peak-voxel level significance ($p_{\text{FWE-voxel}} < .05$). Although the previous contrast between devalued slips and correct upvalued go responses maximizes the difference between habitual versus goal-directed control, the conditions differ in terms of the original training outcome value (as well as test value). To mitigate this, we proceeded to compare devalued slips to still-valuable go responses, which only differ in their test outcome value. Thus, this contrast compares trials on which participants correctly continued responding according to the learned S–R association with trials on which they failed to override this association. Although we have used the same approach previously (in the study of Watson et al.,

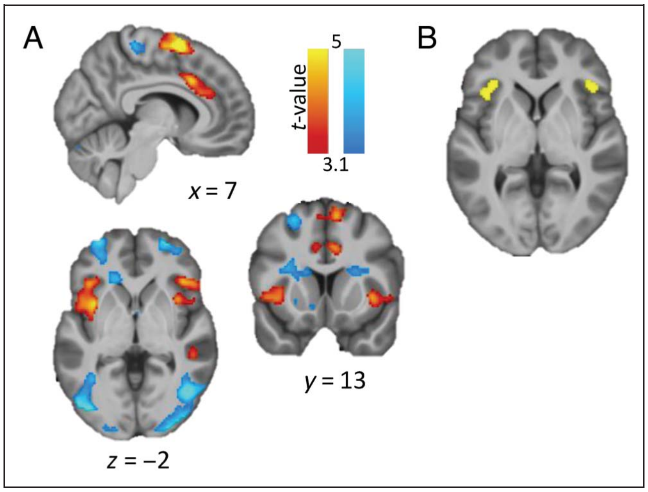


Figure 4. (A) Neural correlates of slips of action in the test phase, as revealed by increased (red – yellow) and decreased (dark – light blue) activity during devalued slips compared with upvalued responses. Clusters that survived whole-brain FWE correction include increased activity in a fronto-parietal network, including the left anterior insula extending to the inferior lateral pFC, SMA, dorsal anterior cingulate cortex, bilateral inferior parietal lobule, and supramarginal gyrus. Conversely, lower activity was seen in the left anterior cingulate cortex extending into caudate nucleus, premotor/primary motor cortex, left lateral OFC, bilateral superior parietal lobe, and several occipital/primary visual areas. Results are shown here at $p < .001$ (uncorrected) for visual purposes, overlaid on the mean T1 image of all participants. (B) The bilateral anterior insula was found to be commonly activated during devalued slips ($x = \pm 40$, $y = 26$, $z = 2$). Shown here in yellow are the voxels that overlap between all four contrasts comparing devalued slips relative to correct (non-)responses during still-valuable, still-not-valuable, devalued and upvalued trials (thresholded at $p < .001$ uncorrected).

2018, the "slips versus respond valuable" contrast), this contrast was not preregistered and should thus be considered exploratory. Similar to the comparison with upvalued go responses, this comparison of slips with still-valuable go responses revealed increased anterior insula activity (bilaterally) during slips, but decreased activity in vMPFC (extending to NAcc), primary motor cortex, paracentral lobule, a large occipital cluster, and large parietal clusters (bilateral) including the angular gyrus and inferior parietal lobule (all $p_{\text{FWE-cluster}} < 0.05$; Figure 4A).

As preregistered, we also compared upvalued no-go responses ("inhibition slips") to correct devalued (no-go) trials, but this did not reveal any significant activation patterns. Moreover, we were not able to conduct the contrast between upvalued and still-valuable no-go trials, because of the low number of omission errors on still-valuable trials.

Our results thus identify the anterior insula as a common region associated with slips toward devalued outcomes, as activity in this region was higher during slips than during go responses toward upvalued and still-valuable outcomes. However, both contrasts are confounded by expected value (the outcome value during the test phase) as they both compare stimuli signaling a nonvaluable outcome (devalued) with stimuli signaling a valuable outcome (upvalued or still-valuable). To control for this, we ran some additional exploratory analyses, comparing activity during devalued slips with correct no-go responses on devalued and still-not-valuable trials. Although these contrasts are difficult to interpret by themselves—they are themselves confounded by pressing a button or not—looking at the overlap between all four contrasts overcomes the value-related confounds and hence could find a common process in the expression of habits. To this end, we used ImCalc to create binary images of all four contrasts thresholded at $t(41) = 3.1$ (equivalent to $p < .001$ uncorrected) and multiply them. The result of this inclusive masking analysis, which is akin to a conjunction analysis, shows that the bilateral anterior insula was commonly activated across all four contrasts (Figure 4B).

Symmetrical Outcome-Revaluation Test: Comparing Goal and Implementation Intentions

None of the planned contrasts comparing test-phase trials trained with implementation with goal intentions revealed significant activation patterns.

DISCUSSION

The aim of the present study was to investigate whether the brain can strategically go on automatic pilot. We investigated this by measuring the impact of strategic planning (i.e., implementation intentions vs. goal intentions) on the acquisition of instrumental actions as well as subsequent flexible, behavioral adjustment. When strategic planning was first introduced during the instrumental learning phase of our paradigm, implementation intentions improved performance relative to goal intentions. Furthermore, in line with the idea that their beneficial effect was mediated by accelerated S–R learning, an exploratory analysis revealed that implementation intentions were associated with reduced activity in the anterior caudate, a brain area previously implicated in goal-directed control (Watson et al., 2018; Liljeholm, Tricomi, O'Doherty, & Balleine, 2011). These effects of strategic planning on performance and neural activity were only apparent early in training, with participants reaching high levels of accuracy (and reduced activity in the anterior caudate) by the end of the learning phase independent of intention type. Our central question, however, was whether implementation intentions would actually impede performance when flexible, behavioral adjustment was required during the subsequent outcome-revaluation test. Importantly, we found no evidence for a detrimental effect of strategic planning on the ability to adapt behavior to changing outcome values, nor any effect on underlying neural activity patterns. We conclude that strategic planning of S–R mappings may allow people to go on automatic pilot to increase behavioral efficiency, but that this does not have to come at the expense of behavioral flexibility. Therefore, mental rehearsal of S–R links does not appear to suffice for the formation of a rigid habit, refuting the notion of "instant habits" and suggesting that behavioral repetition may be crucial for the development of rigid habits.

To shed light on the implications of these findings, we will first discuss them in some more detail, including the basic results (i.e., across strategies) on the relatively novel SORT. First, during the instrumental learning phase, we observed increasing accuracy and decreasing RTs over the course of training, suggesting that participants acquired the S–R mappings. This interpretation was further supported by high levels of subjective automaticity of responding at the end of the instrumental learning phase and increasing involvement of two distinct parts of the dorsal striatum: the posterior putamen, replicating findings from Tricomi et al. (2009), and the caudate nucleus head. Several previous fMRI studies have indirectly implicated the dorsal striatum in habit learning, either showing that with longer instrumental training this region becomes more active (Wunderlich, Dayan, & Dolan, 2012; Tricomi et al., 2009) or that functional connectivity with the (pre-)motor cortex increases (Zwosta et al., 2018; Horga et al., 2015). Although increased activity of the posterior putamen was only significant with small-volume correction (so not very robust), activity of the caudate nucleus as well as the hippocampus survived whole-brain correction. Both regions have previously been implicated in the encoding of S–R representations (McNamee, Liljeholm, Zika, & O'Doherty, 2015). Moreover, we found decreasing activity of the primary motor cortex (extending to mid-posterior cingulate), the hippocampus (extending into the putamen), bilateral temporal cortex, and the right anterior caudate, previously implicated in goal-directed control (Liljeholm et al., 2011; Balleine & O'Doherty, 2010). In line with previous findings, these results suggest that dissociable neural regions support instrumental learning. Notably, these findings were specific to learning to make go responses for valuable outcomes. In contrast, we did not see any changes in neural activity over the course of no-go training, despite high accuracy and reported automaticity at the end of training. Thus, our neuroimaging analyses do not provide evidence for the development of "inhibition habits" (Jahanshahi, Obeso, Rothwell, & Obeso, 2015).

Importantly, when strategic planning was introduced after the first 12 training blocks, implementation intentions initially improved go performance (reflected by higher accuracy during the first block-set than in the preceding (pre-)training block-set), whereas goal intentions impaired it (reflected by lower accuracy and slower RTs). In contrast, no-go learning (i.e., withholding a response for nonvaluable stimuli) was not affected by planning strategy. These findings replicate our previous results (van Timmeren & de Wit, 2022). Furthermore, in line with the notion of "instant habits," this behavioral effect of implementation intentions early in training was accompanied by reduced activity in the anterior caudate relative to goal intentions. This early effect of implementation intentions quickly disappeared, however, and no differences with goal intentions were observed on accuracy and RTs in later training blocks, nor on subjective automaticity after training. Therefore, in support of the notion of strategic automaticity, it appears that instrumental acquisition initially benefitted from if–then planning, while dependency on goal-directed control (as suggested by anterior caudate activity) was reduced.

So far, there have been very few neuroimaging studies that have compared the use of implementation and goal intentions to support behavioral performance. One study (Gilbert, Gollwitzer, Cohen, Oettingen, & Burgess, 2009) showed that implementation intentions engaged the medial BA10 more (and lateral BA10 less) than a control condition, which was argued to reflect increased cue monitoring (and reduced internal information processing). This contrasts with our finding that implementation intentions lead to reduced engagement of the anterior caudate during the instrumental learning phase. However, their control condition was very different to ours. Whereas they specified the cue and the outcome that it signaled to be available (i.e., the S-O contingency; "if the cue appears, then I can score 5 points"), we used a goal intention control condition that specified the R-O contingency, which is arguably more akin to a typical goal intention (e.g., "I will exercise to lose weight").

In the next phase of the SORT, signaled outcome values changed, requiring flexible adaptation of responding to the discriminative stimuli. This allowed us to determine whether strategic planning (during training) would induce the rigidity that is commonly regarded as a hallmark of learned habits that are stamped in through behavioral repetition. However, we failed to find convincing evidence that if–then planning impaired the ability to flexibly adjust responding when signaled outcome values changed. This was despite the fact that participants struggled to adjust learned S–R mappings overall, as reflected in a strong main effect of congruency. Furthermore, in line with the behavioral findings, we also found no evidence for an impact of planning on neural activation patterns during the extinction test phase. Therefore, this first neuroimaging investigation of the effect of implementation intentions on behavioral flexibility in an outcome-revaluation paradigm failed to provide evidence for a shift from goal-directed toward habitual control.

The evidence for intact behavioral flexibility despite if–then planning contrasts with results from an earlier study with this paradigm (van Timmeren & de Wit, 2022). In that study, we found that implementation intentions impaired test-phase performance overall, but this did not lead to inflexibility as would be reflected by lower accuracy on incongruent trials specifically. This general impairment was most likely because of the fact that implementation intentions, by focusing attention on the S–R mappings, blocked learning about the S-O contingencies. To prevent this from happening in the present study, we altered our paradigm to promote active S-O learning at first training phase, before intentions were introduced. As a result, participants already acquired high levels of S-O knowledge when they started using strategic planning. Integrating findings from both studies, it appears that when the agent has full knowledge of the (S-O) contingencies, implementation intentions do not impair flexibility. This finding is encouraging, because in most applied situations in real life, agents are perfectly aware of the three-term instrumental contingencies. Therefore, our results are in line with the idea of implementation intentions being "flexibly tenacious" (Legrand, Bieleke, Gollwitzer, & Mignon, 2017; Gollwitzer, Parks-Stamm, Jaudas, & Sheeran, 2008): People benefit from if–then planning when the situation specified in their plan is encountered (here in terms of higher accuracy and lower RTs during training), but are goal-directed in the sense that they only act on these planned S–R mappings when the signaled outcome is currently a goal.

Across intentions, however, we found that action slips toward devalued outcomes were associated with increased bilateral insula (both when compared with still-valuable and upvalued responses), replicating findings from the only study to date looking at neural activity during slips of action (Watson et al., 2018). The insula is a functionally heterogeneous region (Uddin, Nomi, Hébert-Seropian, Ghaziri, & Boucher, 2017), but the anterior part has been critically implicated in error and salience processing and response inhibition (Uddin, 2015; Chang, Yarkoni, Khaw, & Sanfey, 2013). Specifically, previous work shows that failure to inhibit a learned response (on the stop-signal paradigm) is associated with bilateral insular activity (Ramautar, Slagter, Kok, & Ridderinkhof, 2006). In addition, when compared with upvalued responses, slips were associated with increased activity in the dorsal anterior cingulate cortex, the SMA, and parietal cortex, all part of the salience network (Seeley et al., 2007). Conversely, lower activity during slips was seen in the vMPFC, or medial OFC, when compared with responses for still-valuable outcomes. Previous outcome-devaluation studies suggest that activity in this region mediates goal-directed instrumental learning (de Wit et al., 2012; Valentin et al., 2007). A similar contrast, comparing devalued action slips with responses toward upvalued outcomes, showed lower activity in the lateral OFC and ACC/caudate nucleus head, regions that have also been implicated in goal-directed control (e.g., Watson et al., 2018; McNamee et al., 2015; de Wit et al., 2012). Overall, our results suggest that habitual slips of action arise as a consequence of lapses in goal-directed control (as reflected by decreased activity in these regions) rather than by increased activation of S–R habit regions (i.e., the dorsal striatum). Finally, the informal conjunction analysis of devalued slips (Figure 4B), controlling for differences in expected value and motor response, showed that the anterior insula was commonly activated across all contrasts, implicating it as a key region mediating habitual action slips.

A lack of reliable, positive markers of habits is an important issue in human habit research (Watson, O'Callaghan, Perkes, Bradfield, & Turner, 2022; Kruglanski & Szumowska, 2020; De Houwer, 2019; De Houwer, Tanaka, Moors, & Tibboel, 2018; Watson & de Wit, 2018). In the context of the present study, it begs the question whether habit strength independently contributes to stimulus-dependent, outcome-insensitive responding (i.e., slips of action). A recently published study with the SORT adds weight to this concern, as we showed there that extensive instrumental training failed to impair test performance (Watson, Gladwin, et al., 2022). The lack of reliable evidence for overtraining effects (see also de Wit et al., 2018) could mean different things, but our current findings may offer an interesting explanation. Specifically, we observed that when the planning manipulation was first introduced during training, not only did implementation intentions improve performance, but goal intentions also significantly impaired performance. This may indicate that participants' spontaneous strategy up to that point had not been to form goal intentions, but instead to switch as soon as they could to the more efficient strategy of focusing on the S–R mappings. In other words, they may have spontaneously formed implementation intentions (Bieleke & Keller, 2021). Therefore, rather than improving their performance with the explicit implementation intention manipulation, we impaired it in the goal intention condition. Such an early shift to reliance on S–R associations (i.e., within 12 blocks of training) may explain that previous experimental studies failed to find evidence for overtraining, as their short training conditions may already have been sufficiently long to induce this, and beyond that early shift additional training may not have significantly enhanced the strength of those associations. This idea accords well with results from a study by Pool and colleagues (2022) who found that, following outcome devaluation on a free-operant task, already after moderate training (12 blocks), outcome-insensitive habitual responding was seen in the majority of participants. Our findings further reinforce this interpretation by showing significant changes in neural activity over the course of this relatively short training, with activity of the anterior caudate (implicated in goal-directed learning) decreasing and of the dorsal striatum (implicated in habitual control) increasing. From our study it is unclear, however, how activity in these regions developed in the earliest stages of instrumental training, as that took place outside the scanner. Future research should determine how many behavioral repetitions it takes to permit this shift to an S–R strategy, by assessing the effect of a goal intention manipulation at different time points during training. Our hypothesis is that, at the start of training, this would not yet have a negative impact—relative to implementation intentions—but that it will after a few blocks.

In conclusion, we provide evidence for increased efficiency but preserved flexibility following strategic if–then planning. These behavioral findings were mirrored in our analyses of the underlying brain activity: Implementation intentions did not reduce the engagement of goal-directed control when goals changed, nor increase activity in habit regions. Therefore, our findings suggest that this strategic planning technique supports the implementation of a new target behavior while still allowing for flexible adjustment when goals change.

Reprint requests should be sent to Tim van Timmeren, Department of Social, Health and Organizational Psychology, Utrecht University, Heidelberglaan 1, Utrecht, The Netherlands, 3508 TC, or via e-mail: t.vantimmeren@uu.nl.

Data Availability Statement

Data to recreate the main behavioral analyses (with analysis pipeline and output) are available at OSF: https://www.doi.org/10.17605/OSF.IO/642QU. Whole-brain *t*-maps (without thresholding) of the main fMRI contrasts are available at https://neurovault.org/collections/13191/.

Author Contributions

Tim van Timmeren: Conceptualization; Data curation; Formal Analysis; Investigation; Methodology; Project administration; Visualization; Writing—Original draft; Writing—Review & editing. John O'Doherty: Conceptualization; Writing—Review & editing. Nadza Dzinalija: Investigation; Project administration; Writing—Review & editing. Sanne de Wit: Conceptualization; Funding Acquisition; Resources; Supervision; Writing—Original draft; Writing—Review & editing.

Funding Information

Tim van Timmeren, VIDI grant from Dutch Research Council (Nederlandse Organisatie voor Wetenschappelijk Onderzoek) (https://dx.doi.org/10.13039/501100001722), grant number: Van der Gaag Fund. Sanne de Wit, VIDI grant from Dutch Research Council (Nederlandse Organisatie voor Wetenschappelijk Onderzoek) (https://dx.doi.org/10.13039/501100003246), grant number: 016.145.382.

Diversity in Citation Practices

Retrospective analysis of the citations in every article published in this journal from 2010 to 2021 reveals a persistent pattern of gender imbalance: Although the proportions of authorship teams (categorized by estimated gender identification of first author/last author) publishing in the *Journal of Cognitive Neuroscience* (*JoCN*) during this period were M(an)/M = .407, W(oman)/M = .32, M/W = .115, and W/W = .159, the comparable proportions for the articles that these authorship teams cited were M/M = .549, W/M = .257, M/W = .109, and W/W = .085 (Postle and Fulvio, *JoCN*, 34:1, pp. 1–3). Consequently, *JoCN* encourages all authors to consider gender balance explicitly when selecting which articles to cite and gives them the opportunity to report their article's gender citation balance. The authors of this article report its proportions of citations by gender category to be as follows: M/M = .512, W/M = .256, M/W = .070, and W/W = .163.

REFERENCES

Adams, C. D., & Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. *Quarterly Journal of Experimental Psychology, Section B*, *33*, 109–121. https://doi.org/10.1080/14640748108400816

Balleine, B. W. (2019). The meaning of behavior: Discriminating reflex and volition in the brain. *Neuron*, *104*, 47–62. https://doi.org/10.1016/j.neuron.2019.09.024, PubMed: 31600515

Balleine, B. W., & O'Doherty, J. P. (2010). Human and rodent homologies in action control: Corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*, *35*, 48–69. https://doi.org/10.1038/npp.2009.131, PubMed: 19776734

Bieleke, M., & Keller, L. (2021). Individual differences in if–then planning: Insights from the development and application of the if–then planning scale (ITPS). *Personality and Individual Differences*, *170*, 110500. https://doi.org/10.1016/j.paid.2020.110500

Brandstätter, V., Lengfelder, A., & Gollwitzer, P. M. (2001). Implementation intentions and efficient action initiation. *Journal of Personality and Social Psychology*, *81*, 946–960. https://doi.org/10.1037/0022-3514.81.5.946, PubMed: 11708569

Chang, L. J., Yarkoni, T., Khaw, M. W., & Sanfey, A. G. (2013). Decoding the role of the insula in human cognition: Functional parcellation and large-scale reverse inference. *Cerebral Cortex*, *23*, 739–749. https://doi.org/10.1093/cercor/bhs065, PubMed: 22437053

De Houwer, J. (2019). On how definitions of habits can complicate habit research. *Frontiers in Psychology*, *10*, 2642. https://doi.org/10.3389/fpsyg.2019.02642, PubMed: 31849762

De Houwer, J., Tanaka, A., Moors, A., & Tibboel, H. (2018). Kicking the habit: Why evidence for habits in humans might be overestimated. *Motivation Science*, *4*, 50–59. https://doi.org/10.1037/mot0000065

Deichmann, R., Gottfried, J. A., Hutton, C., & Turner, R. (2003). Optimized EPI for fMRI studies of the orbitofrontal cortex. *Neuroimage*, *19*, 430–441. https://doi.org/10.1016/S1053-8119(03)00073-9, PubMed: 12814592

Delorme, C., Salvador, A., Valabrègue, R., Roze, E., Palminteri, S., Vidailhet, M., et al. (2016). Enhanced habit formation in Gilles de la Tourette syndrome. *Brain*, *139*, 605–615. https://doi.org/10.1093/brain/awv307, PubMed: 26490329

de Wit, S., Corlett, P. R., Aitken, M. R., Dickinson, A., & Fletcher, P. C. (2009). Differential engagement of the ventromedial prefrontal cortex by goal-directed and habitual behavior toward food pictures in humans. *Journal of Neuroscience*, *29*, 11330–11338. https://doi.org/10.1523/JNEUROSCI.1639-09.2009, PubMed: 19741139

de Wit, S., Kindt, M., Knot, S. L., Verhoeven, A. A. C., Robbins, T. W., Gasull-Camos, J., et al. (2018). Shifting the balance between goals and habits: Five failures in experimental habit induction. *Journal of Experimental Psychology: General*, *147*, 1043–1065. https://doi.org/10.1037/xge0000402, PubMed: 29975092

de Wit, S., Niry, D., Wariyar, R., Aitken, M. R. F., & Dickinson, A. (2007). Stimulus-outcome interactions during instrumental discrimination learning by rats and humans. *Journal of Experimental Psychology: Animal Behavior Processes*, *33*, 1–11. https://doi.org/10.1037/0097-7403.33.1.1, PubMed: 17227190

de Wit, S., Watson, P., Harsay, H. A., Cohen, M. X., ven de Vijver, I., & Ridderinkhof, K. R. (2012). Corticostriatal connectivity underlies individual differences in the balance between habitual and goal-directed action control. *Journal of Neuroscience*, *32*, 12066–12075. https://doi.org/10.1523/JNEUROSCI.1088-12.2012, PubMed: 22933790

Dickinson, A. (1985). Actions and habits: The development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, *308*, 67–78. https://doi.org/10.1098/rstb.1985.0010

Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., & Gorgolewski, K. J. (2017). MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PLoS One*, *12*, e0184661. https://doi.org/10.1371/journal.pone.0184661, PubMed: 28945803

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., et al. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*, 111–116. https://doi.org/10.1038/s41592-018-0235-4, PubMed: 30532080

Gardner, B., Abraham, C., Lally, P., & de Bruijn, G.-J. (2012). Towards parsimony in habit measurement: Testing the convergent and predictive validity of an automaticity subscale of the self-report habit index. *International Journal of Behavioral Nutrition and Physical Activity*, *9*, 102. https://doi.org/10.1186/1479-5868-9-102, PubMed: 22935297

Gilbert, S. J., Gollwitzer, P. M., Cohen, A. L., Oettingen, G., & Burgess, P. W. (2009). Separable brain systems supporting cued versus self-initiated realization of delayed intentions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 905–915. https://doi.org/10.1037/a0015535, PubMed: 19586260

Gollwitzer, P. M. (1993). Goal achievement: The role of intentions. *European Review of Social Psychology*, *4*, 141–185. https://doi.org/10.1080/14792779343000059

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist*, *54*, 493–503. https://doi.org/10.1037/0003-066X.54.7.493

Gollwitzer, P. M. (2014). Weakness of the will: Is a quick fix possible? *Motivation and Emotion*, *38*, 305–322. https://doi.org/10.1007/s11031-014-9416-3

Gollwitzer, P. M., Parks-Stamm, E. J., Jaudas, A., & Sheeran, P. (2008). Flexible tenacity in goal pursuit. In J. Y. Shah & W. L. Gardner (Eds.), *Handbook of motivation science* (pp. 325–341). Guilford Press.

Gollwitzer, P. M., & Sheeran, P. (2006). Implementation intentions and goal achievement: A meta-analysis of effects and processes. *Advances in Experimental Social Psychology*, *38*, 69–119. https://doi.org/10.1016/S0065-2601(06)38002-1

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., et al. (2011). Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in Python. *Frontiers in Neuroinformatics*, *5*, 13. https://doi.org/10.3389/fninf.2011.00013, PubMed: 21897815

Hardwick, R. M., Forrence, A. D., Krakauer, J. W., & Haith, A. M. (2019). Time-dependent competition between goal-directed and habitual response preparation. *Nature Human Behaviour*, *3*, 1252–1262. https://doi.org/10.1038/s41562-019-0725-0, PubMed: 31570762

Horga, G., Maia, T. V., Marsh, R., Hao, X., Xu, D., Duan, Y., et al. (2015). Changes in corticostriatal connectivity during reinforcement learning in humans. *Human Brain Mapping*, *36*, 793–803. https://doi.org/10.1002/hbm.22665, PubMed: 25393839

Jahanshahi, M., Obeso, I., Rothwell, J. C., & Obeso, J. A. (2015). A fronto-striato-subthalamic-pallidal network for goal-directed and habitual inhibition. *Nature Reviews Neuroscience*, *16*, 719–732. https://doi.org/10.1038/nrn4038, PubMed: 26530468

JASP Team. (2018). JASP (version 0.8.6) [Computer Software]. Retrieved from https://jasp-stats.org/2018/02/28/now-jasp-0-8-6/.

Kruglanski, A. W., & Szumowska, E. (2020). Habitual behavior is goal-driven. *Perspectives on Psychological Science*, *15*, 1256–1271. https://doi.org/10.1177/1745691620917676, PubMed: 32569529

Legrand, E., Bieleke, M., Gollwitzer, P. M., & Mignon, A. (2017). Nothing will stop me? Flexibly tenacious goal striving with implementation intentions. *Motivation Science*, *3*, 101–118. https://doi.org/10.1037/mot0000050

Liljeholm, M., Dunne, S., & O'Doherty, J. P. (2015). Differentiating neural systems mediating the acquisition vs. expression of goal-directed and habitual behavioral control. *European Journal of Neuroscience*, *41*, 1358–1371. https://doi.org/10.1111/ejn.12897, PubMed: 25892332

Liljeholm, M., Tricomi, E., O'Doherty, J. P., & Balleine, B. W. (2011). Neural correlates of instrumental contingency learning: Differential effects of action-reward conjunction and disjunction. *Journal of Neuroscience*, *31*, 2474–2480. https://doi.org/10.1523/JNEUROSCI.3354-10.2011, PubMed: 21325514

McNamee, D., Liljeholm, M., Zika, O., & O'Doherty, J. P. (2015). Characterizing the associative content of brain structures involved in habitual and goal-directed actions in humans: A multivariate fMRI study. *Journal of Neuroscience*, *35*, 3764–3771. https://doi.org/10.1523/JNEUROSCI.4677-14.2015, PubMed: 25740507

Morris, R. W., Quail, S. L., Griffiths, K. R., Green, M. J., & Balleine, B. W. (2015). Corticostriatal control of goal-directed action is impaired in schizophrenia. *Biological Psychiatry*, *77*, 187–195. https://doi.org/10.1016/j.biopsych.2014.06.005, PubMed: 25062683

Orbell, S., & Verplanken, B. (2010). The automatic component of habit in health behavior: Habit as cue-contingent automaticity. *Health Psychology*, *29*, 374–383. https://doi.org/10.1037/a0019596, PubMed: 20658824

Parks-Stamm, E. J., Gollwitzer, P. M., & Oettingen, G. (2007). Action control by implementation intentions: Effective cue detection and efficient response initiation. *Social Cognition*, *25*, 248–266. https://doi.org/10.1521/soco.2007.25.2.248

Pool, E. R., Gera, R., Fransen, A., Perez, O. D., Cremer, A., Aleksic, M., et al. (2022). Determining the effects of training duration on the behavioral expression of habitual control in humans: A multilaboratory investigation. *Learning & Memory*, *29*, 16–28. https://doi.org/10.1101/lm.053413.121, PubMed: 34911800

Ramautar, J. R., Slagter, H. A., Kok, A., & Ridderinkhof, K. R. (2006). Probability effects in the stop-signal paradigm: The insula and the significance of failed inhibition. *Brain Research*, *1105*, 143–154. https://doi.org/10.1016/j.brainres.2006.02.091, PubMed: 16616048

Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H., et al. (2007). Dissociable intrinsic connectivity networks for salience processing and executive control. *Journal of Neuroscience*, *27*, 2349–2356. https://doi.org/10.1523/JNEUROSCI.5587-06.2007, PubMed: 17329432

Sheeran, P., & Webb, T. L. (2016). The intention–behavior gap. *Social and Personality Psychology Compass*, *10*, 503–518. https://doi.org/10.1111/spc3.12265

Tanaka, S. C., Balleine, B. W., & O'Doherty, J. P. (2008). Calculating consequences: Brain systems that encode the causal effects of actions. *Journal of Neuroscience*, *28*, 6750–6755. https://doi.org/10.1523/JNEUROSCI.1808-08.2008, PubMed: 18579749

Thorndike, E. L. (1911). *Animal intelligence: Experimental studies*. The Macmillan Company. https://doi.org/10.5962/bhl.title.55072

Tricomi, E., Balleine, B. W., & O'Doherty, J. P. (2009). A specific role for posterior dorsolateral striatum in human habit learning. *European Journal of Neuroscience*, *29*, 2225–2232. https://doi.org/10.1111/j.1460-9568.2009.06796.x, PubMed: 19490086

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., et al. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, *15*, 273–289. https://doi.org/10.1006/nimg.2001.0978, PubMed: 11771995

Uddin, L. Q. (2015). Salience processing and insular cortical function and dysfunction. *Nature Reviews Neuroscience*, *16*, 55–61. https://doi.org/10.1038/nrn3857, PubMed: 25406711

Uddin, L. Q., Nomi, J. S., Hébert-Seropian, B., Ghaziri, J., & Boucher, O. (2017). Structure and function of the human insula. *Journal of Clinical Neurophysiology*, *34*, 300–306. https://doi.org/10.1097/WNP.0000000000000377, PubMed: 28644199

Valentin, V. V., Dickinson, A., & O'Doherty, J. P. (2007). Determining the neural substrates of goal-directed learning in the human brain. *Journal of Neuroscience*, *27*, 4019–4026. https://doi.org/10.1523/JNEUROSCI.0564-07.2007, PubMed: 17428979

van Timmeren, T., & de Wit, S. (2022). Instant habits versus flexible tenacity: Do implementation intentions accelerate habit formation? *Quarterly Journal of Experimental Psychology*. https://doi.org/10.1177/17470218221147024, PubMed: 36476147

Watson, P., & de Wit, S. (2018). Current limits of experimental research into habits and future directions. *Current Opinion in Behavioral Sciences*, *20*, 33–39. https://doi.org/10.1016/j.cobeha.2017.09.012

Watson, P., Gladwin, T. E., Verhoeven, A. A. C., & de Wit, S. (2022). Investigating habits in humans with a symmetrical outcome-revaluation task. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01922-4, PubMed: 35867208

Watson, P., O'Callaghan, C., Perkes, I., Bradfield, L., & Turner, K. (2022). Making habits measurable beyond what they are not: A focus on associative dual-process models. *Neuroscience & Biobehavioral Reviews*, *142*, 104869. https://doi.org/10.1016/j.neubiorev.2022.104869, PubMed: 36108980

Watson, P., van Wingen, G., & de Wit, S. (2018). Conflicted between goal-directed and habitual control, an fMRI investigation. *eNeuro*, *5*, ENEURO.0240-18.2018. https://doi.org/10.1523/ENEURO.0240-18.2018, PubMed: 30310863

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*, 291–298. https://doi.org/10.1177/1745691611406923, PubMed: 26168519

Wunderlich, K., Dayan, P., & Dolan, R. J. (2012). Mapping value based planning and extensively trained choice in the human brain. *Nature Neuroscience*, *15*, 786–791. https://doi.org/10.1038/nn.3068, PubMed: 22406551

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European Journal of Neuroscience*, *19*, 181–189. https://doi.org/10.1111/j.1460-9568.2004.03095.x, PubMed: 14750976

Zwosta, K., Ruge, H., Goschke, T., & Wolfensteller, U. (2018). Habit strength is predicted by activity dynamics in goal-directed brain systems during training. *Neuroimage*, *165*, 125–137. https://doi.org/10.1016/j.neuroimage.2017.09.062, PubMed: 28970144