## Participant Details and Data-Exclusion Experiments 1-4

A total of 600, 525, and 420 participants were recruited for participation in Experiments 1, 2, and 3, respectively, from the online participant recruitment platform Prolific Academia (https://www.prolific.ac/). Participants from the earlier experiments were added to a study exclusion list such that only participants who had not participated in any of the previous experiments were able to participate. In Experiment 4, we recruited 184 undergraduate students from Ghent University for participation. The experiment was programmed in Inquisit 4.0 and hosted via Inquisit Web (Millisecond Software, Seattle, WA). The sample size of all experiments was determined on the basis of an a priori power analysis in G*Power 3.1 (Faul, Erdfelder, Lang, & Buchner, 2007) taking into account the results of the previous experiments such that the sample size would provide sufficient power (i.e., power > 0.80) to detect a small effect in Experiment 1 and a small to medium effect in Experiments 2 and 3 (target *d*s = 0.20-0.30). In Experiment 4, target number of participants was 204 which would allow sufficient power (0.80) to observe a moderate difference in outcomes between the two experimental conditions ($d$ = 0.35), which was the minimum effect size observed in Experiment 3. In line with pre-registration, however, we stopped data collection when two months of data collection has passed and only 184 students had participated (due to the limited number of participants of the participant pool).

In line with recommendations by Zhou and Fishbach (2016) to prevent selective attrition, participants were (1) informed about the duration of the experiment and (2) warned that dropping out could affect the quality of data and that it is essential for scientific advance to have good data. Overall dropout rate was 2.8% (Experiment 1), 1.7% (Experiment 2), 4.5% (Experiment 3), and 0.0% (Experiment 4). The dropout rates were comparable across the different task conditions in

all experiments, $\chi^2$s <1.08, *ps* > .34. Hence, there was no evidence for condition-dependent attrition.

Following our preregistered data-analysis plan and standard procedures for data-reduction of implicit evaluation tasks (e.g., Smith, De Houwer, & Nosek, 2013), we excluded the data from participants who (1) had error rates above 30% when considering all IAT test blocks or above 40% for any one of these blocks (Experiment 1: 2.8%; Experiment 2: 8.4%) or (2) completed more than 10% of IAT trials faster than 400ms (Experiment 1: 7.8%; Experiment 2: 2.9%; Experiment 3: 2.9%; Experiment 4: 0.0%). Analyses were performed on the data of 519 participants in Experiment 1 (300 women; mean age = 34, *SD* = 12, range = 18-69; country of residence: 50.9% UK, 41.6% US, 5.1% Canada, 2.3% other), 455 participants in Experiment 2 (288 women, mean age = 34, *SD* = 12, range = 18-69; country of residence: 59.6% UK, 35.4% US, 2.9% Canada, 2.1% other), 389 participants in Experiment 3 (219 women, mean age = 34, *SD* = 13, range = 18-68; country of residence: 52.7% UK, 40.9% US, 3.5% Canada, 2.9% other), and 184 participants in Experiment 4 (159 women, mean age = 20, *SD* = 2, range = 18-33; country of residence: Belgium; mean weight = 63 kg, *SD* = 10, range = 45-98; mean length = 170 cm, *SD* = 8, range = 151-193; mean BMI = 21.6, *SD* = 2.7, range = 16.9-33.1; 6.5% BMI<18.5; 8.7% BMI>25). In Experiment 3, a total of 307 participants completed the second phase of the experiment (78.9% of the participants of Experiment 3). Drop-out rate between the two phases was comparable across the three task conditions of Experiment 3, $\chi^2$s < 0.85, *ps* > .46. Participants in Experiments 1-3 received a monetary reward of 1.50 Great-British Pounds (with Experiment 3 participants receiving an extra 0.50 Great-British Pounds if they completed phase 2). Participants in Experiment 4 received 5 euros for participation (as well as a banana or waffle, see p.16).

## Detailed Procedure Experiments 1-4

### Procedure of Experiment 1

**Introduction.** After participants had given informed consent, they provided answers to demographic questions asking for their age, gender, and country of residence. Participants were then informed that they would learn about two new food brands (i.e., Vekte and Empeya) that would soon be introduced into supermarkets around the world. The brand names and logos were presented below these instructions. The brand names and logos were taken from a recent study investigating approach-avoidance (AA) training effects on evaluations of novel food brands (Hughes, Van Dessel, & Smith, 2017). Brand names and brand logos were pre-rated in an on-line study as evaluatively neutral (i.e., ratings near the mid-point of the 9-point Likert scale ranging from 1[very negative] to 9 [very positive]).

**AA training Task.** Participants were randomly assigned to one of three different versions of the AA training task: two typical AA training tasks (manikin task, avatar task) and one consequence-based AA training task (avatar consequences task). Randomization was established via the randomization algorithm on Inquisit Web.

***Manikin task.*** The AA training manikin task was adapted from Woud, Maas, Becker, and Rinck (2013). We use the manikin task as the reference task for typical AA training (to allow comparison between typical and consequence-based AA training). AA training can be defined as the repeated performance of responses that alter the difference between a specific stimulus and the self (see Krieglmeyer, De Houwer, & Deutsch, 2011; Krieglmeyer, Deutsch, De Houwer, & De Raedt, 2010). The manikin task uses a tried-and-tested procedure that was used in many previous studies on approach-avoidance and has produced strong effects (see Van Dessel, Eder, & Hughes, in press). The task involves movements of a symbolic representation of the self rather

than movements that involve a distance change between a stimulus and (parts of) the actual self (e.g., joystick movements). Early accounts of AA training effects assumed that AA behavior activates motivational systems of AA via the proprioceptive feedback that this behavior produces (Cacioppo, Priester, & Berntson, 1993) and hence many AA training studies used joystick movements for examining AA training effects. However, this strong embodiment view (i.e., the view that embodiment effects require actual proprioceptive feedback from motor behavior) has been largely abandoned (see Neumann & Strack, 2000). It is now widely accepted that a symbolic representation of the self is a key component of AA motivation (see Markman & Brendl, 2005). In line with this idea, there is evidence that the manikin procedure triggers spontaneous AA tendencies (Krieglmeyer, Deutsch, De Houwer, & De Raedt, 2010). Furthermore, the manikin task was shown to produce stronger effects than many other variants measuring AA training tendencies such as a joystick task or joystick feedback task (see Krieglmeyer & Deutsch, 2010). Also, in a recent study we found stronger AA training effects on evaluations of both neutral and valenced stimuli with a manikin task than with a joystick task or joystick feedback task (Van Dessel, De Houwer, Gast, & Spruyt, 2017). Hence, we believed the manikin task to be one of the best AA training procedures for producing AA training effects and would serve as a good comparison to compare the effectiveness of current, typical, AA training and consequence-based AA training.

Participants who were assigned to the manikin task were informed that in the first part of the experiment they would see a stick figure (manikin) and products from the two brands. They were then told that their task would be to move the stick figure towards or away from the food products. The instructions then asked participants to memorize the following guidelines. First, participants should approach a food product with a GREEN frame around it by moving the stick

figure towards it and avoid a product with a BLUE frame around it by moving the stick figure away from it. Second, participants should imagine that the manikin represented them and that when they moved the manikin up or down it was them who were moving. Participants then proceeded to the next page in which they were asked to indicate which action they would have to perform in response to food products with either a blue frame or a green frame in the following task (response options: 'approach', 'avoid', or 'I don't remember'). Participants proceeded to the task if they were able to indicate the correct contingencies. Elsewise, participants saw the instruction screen again.

The manikin task consisted of 80 trials. Each trial started with the presentation of the manikin, followed by the presentation of a rectangular shape with the brand name logo on it (i.e., the food brand product) after 500ms and a green or blue frame surrounding the product after an additional 250ms (Figure 1). The 250ms delay was installed to facilitate prediction of the approach-avoidance action on the basis of the word identity (which might facilitate contingency awareness, an important moderator of AA training effects: see Van Dessel, De Houwer, & Gast, 2016). One of the brands (e.g., Vekte) was always surrounded by a blue frame and the other brand (e.g., Empeya) was always surrounded by a green frame. Assignment of the brands to the frame colors was counterbalanced across participants. Whenever the participant made a correct response by pressing the up or down key on the keyboard, the manikin moved up or down on the screen. Duration of the manikin movement was 1000ms. An incorrect response produced no error feedback. Participants had to make a correct response on each trial to complete the task. The next trial started 200ms later.

***Avatar task.*** The avatar task was designed for the purpose of this study and served as a typical AA training control for the avatar consequences task described below. Participants

assigned to the avatar task received the same instructions as participants in the manikin task with two exceptions. First, they were told that they would see a fridge with food products from the two brands and that on each trial the fridge door would open and a food product would appear. Second, they were informed that they would also see an avatar and that they had to move the avatar towards or away from the food product by pushing the up or down key (depending on the color of the frame surrounding the product). Before starting the task, participants selected which avatar would represent them in the avatar task by choosing between a male avatar and a woman avatar. Participants then indicated the instructed contingencies between approach-avoidance actions and frame colors.

Each trial of the avatar task, participants saw a virtual kitchen environment with a fridge (programmed in iClone v6.52, Reallusion). The trial started with the presentation of the self-selected avatar. Next, the fridge gradually opened until a food brand product appeared (after 500ms). A green or blue frame was presented surrounding the product after an additional 250ms (Figure 1). Whenever the participant made a correct response by pressing the up or down key on the keyboard, the avatar either moved towards the brand product (this movement ended with the manikin holding the product) or away from the brand product (this movement ended with a refraining hand gesture). Duration of the avatar movement was 1000ms. The next trial started 200ms later.

***Avatar consequences task.*** Compared to participants in the avatar task condition, participants in the avatar consequences task condition received the additional instruction to pay careful attention to the consequences of approaching or avoiding the food products. On each trial of the task, these participants not only saw the avatar, the fridge, and the food product with the frame, but they also saw a health bar presented above the avatar (Figure 1). The health bar and avatar had 7 health level versions (1: health bar empty, avatar very sick; 7: health bar filled,

avatar very healthy). The task started with the avatar at average health level (level 4). After participant made a correct button press response (depending on the colored frame) and the subsequent avatar movement, an action consequence followed such that either (1) the health bar gradually depleted, the sentence 'I feel sick' appeared, and the avatar had a more unhealthy appearance (negative consequence) or (2) the health bar filled, the sentence 'I feel healthy' appeared, and the avatar had a more healthy appearance (positive consequence). Crucially, in this task, products from both brands were presented equally often with blue and green frames (and were approached and avoided an equal number of times). However, approaching always led to positive consequences and avoidance to negative consequences for one brand, whereas approaching always led to negative consequences and avoidance to positive consequences for the other brand. Duration of the consequence-related feedback was 500ms. The next trial started 200ms later.

**Outcome measures.** ***Consumer choice question.*** After the AA training, participants completed a question that measured consumer choice. This outcome measure was presented first because (1) we considered this most important as it most strongly (but not perfectly) relates to real-world behavior (Morwitz, 2001) and (2) we wanted to prevent unwanted influences that might result from the performance of the other outcome measurement tasks. Before completing this question, participants were reminded that the two brands would soon introduce food products into supermarkets around the world. They were then asked which of the two food products they would prefer for us to send them a sample of. Participants answered by selecting one of four options of a dropdown menu with "I would try Vekte", "I would try Empeya", "I would try neither", and "I would try both" as possible answers.

***Implicit evaluation task: IAT.*** Following the behavioral intention question, participants performed an IAT that measured implicit evaluations of the two brand names and an explicit rating task that measured explicit evaluations of the brand names. The order of IAT and rating task was counterbalanced. The IAT of Experiment 1 followed the procedure described by Van Dessel, Gawronski, Smith, and De Houwer (2017). In the IAT, participants categorized eight attribute words (e.g., wonderful, evil) as 'positive' or 'negative' and the two brand logos as their respective names ('Vekte' and 'Empeya'). To avoid that the target stimuli were classified only on the basis of simple perceptual features, the brand names were presented in different font types (Arial Black and Fixedsys) and capitalizations (uppercase and lowercase) resulting in 4 different stimuli for each brand. This procedure is often used for the measurement of implicit evaluations of novel stimuli with an IAT (see Zanon, De Houwer, Gast, & Smith, 2014). The attribute words were always presented in Arial Black, font size 16, lower case. The IATs consisted of three practice blocks and two experimental blocks. Participants began with 20 practice trials sorting the brand logos and 20 practice trials sorting positive and negative stimuli. Next, participants completed 56 trials in which Vekte and positive stimuli shared one response key and Empeya and negative stimuli shared another response key (half of the participants completed the IAT in this way, whereas the other participants began by sorting Empeya and positive with the same key). IAT block order was orthogonally crossed with the between-subjects factors of Positive stimulus (whether Vekte or Empeya was the target of positive attitude induction) and AA training task condition. Participants then practiced sorting brand logos on 40 trials with a reversed response key assignment. Finally, participants completed a second set of 56 trials in which one brand shared a response key with negative and the other nonword brand shared a response key with positive. On each trial, a word or brand logo was presented in the center of the screen until the participant pressed one of the two valid keys (i.e., 'F' or 'J'). If the response was correct, the

stimulus disappeared and the next stimulus was presented 400ms later. If the response was incorrect, the stimulus was replaced by a red 'X' for 200ms. The next stimulus appeared 400ms after the red 'X' was removed from the screen.

***Explicit evaluation task: evaluative rating task*.** In the evaluative rating task, participants were asked to rate how positive or negative they found each of the two brands. Participants gave their ratings on a Likert scale ranging from 1 (*very negative*) to 9 (*very positive*). The order of the two questions was randomized.

***Approach-avoidance behavior measurement task.*** Participants were told that they would perform a final task in which they would again see the food products and the manikin (or avatar). Participants received instructions to imagine that they were now at home and that they were free to choose which action to make when seeing the two food products (i.e., approach or avoid). Participants then completed 10 trials of the same AA training task they had completed before but now there were no colored frames surrounding the food products and participants were free to perform whatever action they wanted. There were no consequences attached to the approach or avoidance actions (except that the manikin or avatar moved towards or away from the product).

**Contingency questions.** Participants were asked to complete a manipulation check for each brand. They first answered how often they had approached or avoided the food product in the task at the beginning of the experiment by selecting an option from a dropdown menu with "always", "never", "approximately 50% of the times", and "I didn't notice" as possible answers. Participants in the avatar consequences task condition then answered a follow-up question that asked what happened immediately after they approached or avoided the food product in that task. Possible answers were "the person became more healthy", "the person became less healthy", "the person became more healthy approximately 50% of the times", "nothing happened", and "I didn't notice".

**Additional questions.** Participants then answered five additional questions: (1) They indicated to what extent they had imagined that *they* were the manikin (or avatar) approaching or avoiding the food products on a scale ranging from 1 (*not at all*) to 9 (*very much*). (2) They rated how positive or negative they considered the action of approaching things, in general, on a scale ranging from 1 (*extremely negative*) to 9 (*extremely positive*). (3) They rated how positive or negative they considered the action of avoiding things, in general, on a scale ranging from 1 (*extremely negative*) to 9 (*extremely positive*). (4) They indicated on what basis they had given the evaluative ratings for the nonwords. Possible answers were: "I responded based on what I thought the experimenter wanted me to do", "I responded based on what I learned about the stimuli earlier in the experiment", "I responded based on my feelings about the stimuli", and "I do not know why I gave ratings the way I did". We included this final question to control for demand characteristics that might bias the observed effects (see Sharpe & Whelton, 2016). (5) They indicated whether they thought we tried to influence how they felt about the two products in the AA training task. If they answered yes, a follow-up question asked to what extent they resisted our attempts to influence how they felt on a scale ranging from 1 (*I did not resist your attempts to influence how I felt at all*) to 9 (*I very strongly resisted your attempts to influence how I felt*). Finally, participants were debriefed and thanked for their participation.

**Procedure of Experiment 2**

Experiment 2 was a replication of Experiment 1 with three exceptions. First, participants either performed the manikin task or the avatar consequences task of Experiment 1 or a third task (i.e., a *goal-relevant avatar consequences task*). This third task was included because we expected that this would further enhance consequence-based AA training effects (given that in this task learning the action consequences is relevant for participants' task performance).

Compared to participants in the avatar consequences task condition, participants in the goal-relevant avatar consequences task condition did not receive any information about colored frames surrounding the food products (and color-dependent movements). Rather, they were told that on each trial they would be able to *choose* to either approach a food product by moving the avatar towards it or avoid a food product by moving the avatar away from it. The instructions also informed participants that each time they would perform one of the actions they would see the avatar become more healthy or more sick and that their task would be to make the avatar as healthy as possible by performing the approach and avoidance actions. During the task, there were no colored frames surrounding the brand products and participants could make any action they wanted in response to each of the brand products. Similar to the avatar consequences task, approaching always led to positive consequences and avoidance to negative consequences for one brand, whereas approaching always led to negative consequences and avoidance to positive consequences for the other brand (Figure 1). This task started with a first block of 28 trials followed by two blocks of 26 trials. In between blocks, participants were informed about the health level that they had reached after the previous task block. They were then asked to imagine that time had passed and that they had now become sick again. The next block started with the avatar and health bar presented at the lowest health level.

Second, the IAT in Experiment 2 was a personalized IAT that was identical in measurement procedures to the IAT used in Experiment 1, except that, in accordance with standard recommendations for the personalized IAT (see Han, Olson, & Fazio, 2006), (1) the category labels for the attribute words (‘positive’ and ‘negative’) were changed to ‘I like’ and ‘I dislike’, (2) there was no error feedback, and (3) attribute stimuli were six positive and six negative nouns (e.g., ‘holiday’, ‘accident’: stimuli taken from Haynes, Kemps, & Moffitt, 2016).

Third, prior to the consumption behavior question we first informed participants that,

when providing their payment via the Prolific Academic website, we could also provide a coupon with which they would be able to obtain a free sample of one of the brands when it becomes available. We then asked participants which food product they wanted us to send them a coupon for (possible options: “send me a coupon for Vekte”, “send me a coupon for Empeya”, “send me a coupon for neither”, and “send me a coupon for both”). The adaptation of this behavioral question compared to Experiment 1 was done to improve believability of the question premise.

**Procedure of Experiment 3**

**Phase 1.** Experiment 3 followed the procedure of Experiment 2 with the following exceptions. First, in the introduction, participants received the information that in this study they would see different food items consisting of foods that are available in supermarkets around the world. Stimuli used throughout the experiment always consisted of healthy and unhealthy foods rather than novel food brands (see below for details). At this time, participants were also informed that the study consisted of two phases and that they would be asked to come back for the second phase after one day. We recommended participants to only start the experiment if they would be able to fulfill this requirement.

Second, four questions were asked to participants at the beginning of the experiment (after completion of the demographic questions). Participants indicated (1) to what extent they had the goal to eat healthily (scale ranging from 1 [*not at all*] to 9 [*very much*]), (2) how hungry they felt at that moment (scale ranging from 1 [*not at all*] to 9 [*very much*]), (3) in general, how often they ate healthy (scale ranging from 1 [*I never eat healthy*] to 9 [*I always eat healthy*]), and (4) to what extent they found it difficult to cut down or stop eating unhealthy foods (scale ranging from 1 [*not at all*] to 9 [*very much*]).

Third, the AA training consisted of two versions of the manikin task that had the same procedure as in Experiment 2 with the exception that healthy and unhealthy foods were used as stimuli rather than products from the brands Vekte and Empeya. In one condition, participants always approached healthy foods and avoided unhealthy foods on the basis of the colored frames by moving the manikin *(manikin task condition*). In the other condition, participants approached and avoided healthy and unhealthy foods 50% of the times by moving the manikin (*control condition*: see also Becker et al., 2015). This control condition was added to gain an estimate of the extent to which (typical and consequence-based) AA training can produce shifts in important outcomes that are due to the manipulated (stimulus-response or stimulus-response-consequence) contingencies. For this reason, the control condition matches the stimulus exposure and response requirements of the other conditions but not the contingencies. Note that the 50% control AA training might already involve some change in response bias and might therefore be considered a rather conservative baseline (see Kakosche, Kemps, & Tiggeman, 2018).

The third AA training task condition consisted of the goal-relevant avatar consequences task of Experiment 2 with healthy and unhealthy foods as stimuli. In this task, participants were free to choose their approach-avoidance responses to the food stimuli. Performing the approach action in response to healthy foods always led to positive health outcomes (i.e., the health bar filled up and the avatar looked more healthy) and performing the avoidance action in response to healthy foods led to negative health outcomes (i.e., the health bar depleted and the avatar looked less healthy). For unhealthy foods, an approach action always led to negative health outcomes and an avoidance action led to positive health outcome. The stimuli used in the AA training tasks were 20 standardized food images taken from the Full4Health Image Collection (Charbonnier, van Meer, van der Laan, Viergever, & Smeets, 2016). We selected the 10 items that (1) had high

ratings for recognizability (*M* = 96.4%) and (2) had high healthiness ratings (*M* = 8.33, *SD* = 0.79) (healthy foods), as well as the 10 items that (1) had high ratings for recognizability (*M* = 99.4%) and (2) had low healthiness ratings (*M* = 1.62, *SD* = 0.82) (unhealthy foods).

Fourth, the measure of consumption behavior was a food choice task (Salmon et al. 2014) that consisted of 6 trials in which participants were shown a healthy and an unhealthy food product and had to indicate which product they would prefer to receive a coupon for that they could use in their supermarket. Trials either involved trade-off pairs (i.e., pairs for which the healthy food is less attractive than the unhealthy food) or controlled pairs (i.e. pairs for which healthy and unhealthy foods are equally attractive). Stimuli were depictions of the same food products used by Salmon et al. (2014). The pictures were selected from the food-pics database (Blechert, Meule, Busch, & Ohla, 2014) depending on attractiveness ratings such that the trade-off pairs significantly differed in attractiveness ratings (healthy foods: *M* = 18.3, *SD* = 4.59, unhealthy foods: *M* = 39.4, *SD* = 1.39), *p*s < .001, whereas the controlled pairs did not (healthy foods: *M* = 37.1, *SD* = 6.75, unhealthy foods: *M* = 37.5, *SD* = 8.44), *p*s > .56.

Fifth, in Experiment 3 a personalized IAT was used to measure implicit evaluations of healthy and unhealthy foods. The attribute stimuli consisted of pictures of six healthy and unhealthy food items (3 of which pictures of foods that had also been presented during the AA training phase). These pictures depicted the same stimuli as in the IAT used by Mai, Hoffmann, Hoppert, Schwarz, and Rohm (2015) who found a strong relation between implicit evaluations measured with the IAT and healthy eating behavior.

Sixth, in the evaluative rating task participants were asked how positive or negative they considered healthy or unhealthy foods in general. They provided their ratings on a Likert scale ranging from 1 (*very negative*) to 9 (*very positive*).

Seventh, the approach-avoidance behavior measurement task consisted of 12 AA training

task trials in which participants were free to approach or avoid healthy and unhealthy foods. Six trials involved healthy and unhealthy items that were also used in the previously performed AA training task, the other six trials involved other pictures of healthy and unhealthy foods (also taken from the food-pics database) that had not been used in the experiment.

**Phase 2.** Approximately one day after completing the first phase of the experiment, participants were contacted via the Prolific Academic website and asked to perform the final part of the study. Participants who came back for the second phase of the experiment completed five questions. They indicated (1) how often they had eaten healthy meals or foods in the period between completing the first part of the study and completing these final questions (scale ranging from 1 [*I never ate healthy*] to 9 [*I always ate healthy*]), (2) how often they head eaten unhealthy meals or foods in the period between completing the first part of the study and completing these final questions (scale ranging from 1 [*I never ate unhealthy*] to 9 [*I always ate unhealthy*]), (3) how often they had had the intention to eat healthy in the period between completing the first part of the study and completing these final questions (scale ranging from 1 [*I never had the intention to eat healthy*] to 9 [*I always had the intention to eat healthy*]), (4) to what extent they had found it difficult to refrain from eating unhealthy foods in the period between completing the first part of the study and completing these final questions (scale ranging from 1 [*not at all*] to 9 [*very much*]), and (5) to what extent they now had the goal to eat healthily (scale ranging from 1 [*not at all*] to 9 [*very much*]).

**Procedure of Experiment 4**

Experiment 4 followed the procedure of Experiment 3 with the following exceptions. First the experiment was performed in the lab. Participants were recruited via the University participation website of Sona systems were they were instructed not to eat anything during the

two hours prior to participation (participants were reminded of this 1 day before the scheduled session). Participation could only be scheduled between the hours of 2PM-5PM, during a normal period that individuals would consume additional calories outside of meal time (see Epstein, Dearing, & Roba 2010). Upon participation, participants were picked up by a female experimenter who was a paid Ghent University student. The experimenter was a unaware of the experimenter hypotheses (e.g., unfamiliar with approach-avoidance training) and of the participant task condition. Participants were seated at a laptop computer (Dell) with a 17" CRT monitor attached to it. Participants were then asked to read the instructions on the computer screen and complete the experiment. The experimenter then left the room and participants started by rating healthy eating intentions, hunger, healthy eating behavior and healthy eating behavior difficulty (see Experiment 3).

Second, the AA training consisted only of two versions of the manikin task: control condition the goal-relevant avatar consequences task condition of Experiment 3 with healthy and unhealthy foods.

Third, after performing the AA training, participants completed an ad-libitum snack task, adapted from Haynes, Kemps, & Moffit (2015). Participants were presented with four full bowls of pre-weighed popular energy-dense snack foods: mixed candy (3.3 kcal/g), potato chips (5.3 kcal/g), M&Ms (4.7 kcal/g), and cheese flips (5.1 kcal/g). Participants were informed that they had to rate the foods on different sensory characteristics for use in an unrelated study. They were informed that they could consume as much or as little of the food as they wanted, as long as they at least tasted each food so that they would be able accurately rate the food's characteristics. The four foods were rated on 6 sensory properties (e.g., sweetness, bitterness) using 9-point Likert-type scales. After providing the ratings, participants were informed that they could sample more of the products at liberty.

Fourth, the measure of consumption behavior was the same food choice task as in Experiment 3 but this time participants were informed that they would receive one of their choices after completion of the experiment.

Fifth, after the food choice task, personalized IAT and evaluative rating task, participants completed the final questions that now additionally involved (1) a question asking to what extent the participant was tempted to eat the foods presented in the taste-test (7-point Likert scale: 1 = not at all; 7 = extremely) and (2) an assessment of task engagement based on the short version of the Dundee Stress State Questionnaire (Helton & Näswall, 2015). Specifically, participants rated the extent to which they felt bored and alert in the AA training task as well as the extent to which their attention was directed towards the task and the extent to which they were determined to perform the task well (5-point Likert scale: 1 = not at all; 5 = completely).

Finally, participants' weight and height are measured (for BMI calculation), the participant was paid, given a banana or a waffle (depending on choices in the behavioral intention task) and thanked for participation. The experimenter then measured the weight of each of the four bowls of food.

## Detailed Results Experiments 1-4

### Results of Experiment 1

**AA training task.** We examined participants' mean latencies and errors in the AA training task. Between-subjects *t*-tests revealed that participants were faster in the manikin task ($M$ = 520ms, $SD$ = 148ms) and in the avatar task ($M$ = 490ms, $SD$ = 152ms) than in the avatar consequences task ($M$ = 668ms, $SD$ = 155ms), $t$s > 9.11, $p$s < .001. Participants also made fewer errors in the manikin task ($M$ = 2.47%, $SD$ = 3.15%) and in the avatar task ($M$ = 2.05%, $SD$ = 4.34%) than in the avatar consequences task ($M$ = 3.72%, $SD$ = 5.31%), $t$s > 2.67, $p$s < .008.

**Outcome measure 1: Consumer choice.** A behavioral choice score was computed by recoding participants' responses to the consumption behavior question such that -1 indicates that participants selected only the negative brand (i.e., the brand that was consistently avoided in the manikin or avatar task and the brand for which approach led to negative and avoidance to positive consequences in the consequences tasks), 0 indicates that participants selected both or neither brand, and 1 indicates that participants would try only the positive brand (i.e., the brand that was consistently approached in the manikin or avatar task and the brand for which approach led to positive and avoidance to negative consequences in the consequences tasks). These scores were subjected to a Kruskal-Wallis test, which revealed a significant main effect of AA Task Condition, $\chi^2(2) = 14.66$, $p < .001$. Multiple comparisons were performed using the Mann-Whitney U test and indicated that the scores were higher for the avatar consequences task condition ($M = 0.50$, $SD = 0.68$) than the scores for avatar task ($M = 0.31$, $SD = 0.56$) or manikin task conditions ($M = 0.33$, $SD = 0.60$), $p$s <.003. The scores for the avatar and manikin task conditions were not significantly different, $p = .64$. Overall, participants in all task conditions selected the positive brand more than the negative brand (manikin: $p < .001$, $d_z = 0.54$, avatar: $p < .001$, $d_z = 0.54$, avatar consequences: $p < .001$, $d_z = 0.73$). A Bayes factor, calculated in accordance with Rouder, Speckman, Sun, Morey, and Iverson (2009), indicates that the data provided strong evidence for this effect in all task conditions, $BF_1$s > 10000.

**Outcome measure 2: IAT scores.** The IAT score was calculated on the basis of participants' IAT performance using the D4-algorithm (Greenwald, Nosek, & Banaji, 2003) so that it indicates a preference for the positive brand over the negative brand. Spearman-Brown corrected split-half reliability of this score was $r(517) = .88$. We performed an ANOVA on IAT scores with Task, Positive Brand (Vekte, Empeya), and IAT Block Order (positive brand

stimulus and positive words categorized with the same key first, negative brand stimulus and positive words categorized with the same key first) as between-subjects factors. The main effect of IAT Order was significant, $F(1,507) = 86.04$, $p < .001$, but not the main effect of AA Task Condition, $F(2,507) = 2.07$, $p = .13$. Overall, participants in all task conditions exhibited an implicit preference for the positive brand, all $p$s $< .001$, $BF_1$s $> 10000$ (manikin: $d_z = 0.62$; avatar: $d_z = 0.59$, avatar consequences: $d_z = 0.42$).

**Outcome measure 3: Explicit rating scores.** An explicit rating score was computed by subtracting the liking ratings for the positive brand from the liking ratings for the negative brand. We performed an ANOVA on this rating score with AA Task Condition and Positive Brand as between-subjects factors. The main effect of AA Task Condition was significant, $F(2,513) = 11.61$, $p < .001$. Contrasts revealed that the score was higher for the avatar consequences task condition ($M = 3.32$, $SD = 3.75$) than for the avatar ($M = 1.99$, $SD = 3.03$) or manikin task conditions ($M = 1.76$, $SD = 2.90$), $t$s $> 3.79$, $p$s $< .001$. The scores were not significantly different for avatar and manikin task conditions, $t(341) = 0.66$, $p = .51$. Overall, participants in all task conditions indicated that they preferred the positive brand, all $p$s $< .001$, $BF_1$s $> 10000$ (manikin: $d_z = 0.61$; avatar: $d_z = 0.66$, avatar consequences: $d_z = 0.88$).

**Outcome measure 4: Approach-avoidance scores.** An approach-avoidance score was computed by subtracting (1) the total number of times that participants chose to approach the positive brand or avoid the negative brand from (2) the total number of times that participants chose to avoid the positive brand or approach the negative brand in the free choice AA task. Spearman-Brown corrected split-half reliability of this score was $r(517) = .92$. We performed an ANOVA on this score with AA Task Condition and Positive Brand as between-subjects factors. Only the main effect of AA Task Condition was significant, $F(2,513) = 3.02$, $p = .049$. Contrasts

revealed that the score was higher for participants in the avatar consequences task condition ($M$ = 2.29, $SD$ =3.21) than for participants in the avatar task ($M$ = 1.71, $SD$ = 2.90) or manikin task ($M$ = 1.55, $SD$ =2.80) conditions, $t$s > 1.81, $p$s < .070. The scores were not significantly different for participants in the avatar and manikin task conditions, $t(341) = 0.51$, $p$ =.61. Overall, participants in all task conditions approached the positive brand more than the negative brand, all $p$s <.001, $BF_1$s > 10000 (manikin: $d_z = 0.55$, avatar: $d_z = 0.59$, avatar consequences: $d_z = 0.73$).

**Reactance scores.** An ANOVA on reactance scores also revealed a main effect of AA Task Condition, $F(2,480) = 4.46$, $p = .012$. Participants were more reactant in the manikin task condition ($M$ = 4.23, $SD$ = 2.65) than in the avatar consequences task condition ($M$ = 3.41, $SD$ = 2.33), $t(318) = 2.96$, $p = .003$ (Table 1). Reactance scores were not significantly different for participants in the avatar task ($M$ = 3.86, $SD$ = 2.44) compared to participants in manikin task or avatar consequences task conditions, $t$s < 1.65, $p$s > .10. Reactance scores correlated negatively with explicit rating scores, $r(481) = -.27$, $p < .001$, but not with IAT scores, $r(481) = -.04$, $p = .34$. Correlation coefficients for reactance and explicit rating scores did not differ significantly between task conditions, $p$s > .22. Correlation coefficients for reactance and IAT scores did differ significantly but only between manikin task condition ($r[162] = -.19$) and avatar consequences condition ($r[157] = .08$), $p = .018$.

**Demand.** A Chi-square proportion test indicated that a significantly higher proportion of participants in the manikin task (9.3%) and the avatar task (8.8%) conditions indicated that they had complied to experimental demands when providing their evaluative ratings than in the avatar consequences task condition (2.9%), $\chi^2$s > 4.52, $p$s < .034 (Table 1). Demand compliant participants had higher explicit rating scores ($M$ = 3.31) and IAT scores ($M$ = 0.25) than non-demand compliant participants (explicit rating: $M$ = 1.55, IAT: $M$ = -0.01) in the manikin task

condition, *F*s > 4.07, *p*s < .046, but not in the avatar task condition, *F*s < 1.02, *p*s > .31 (explicit: demand compliant: *M* = 1.54, non-demand compliant: *M* = 2.10; IAT: demand compliant: *M* = -0.11, non-demand compliant: *M* = 0.04), or the avatar consequences task condition, *F*s < 0.34, *p*s > .56 (explicit: demand compliant: *M* = 0.55, non-demand compliant: *M* = 3.53; IAT: demand compliant: *M* = 0.20, non-demand compliant: *M* = 0.07). Excluding the data of demand compliant participants did not change the statistical significance of any of the reported effects.

**Contingency awareness.** A Chi-square proportion test indicated that a significantly higher proportion of participants in the manikin task (76.2%) and avatar task (73.1%) conditions indicated correct contingencies (between stimuli and approach/avoidance actions) than participants in the avatar consequences task condition (55.3%) indicated correct contingencies (between stimuli, approach-avoidance actions and action consequences), $\chi^2$s > 10.21, *ps* < .002.

**Results of Experiment 2**

**AA Task.** We examined participants' mean latencies and errors in the AA task. Between-subjects *t*-tests revealed that participants were faster in the manikin task (*M* = 499ms, *SD* = 154ms) than in the avatar consequences task (*M* = 672ms, *SD* = 167ms) or the goal-relevant avatar consequences task (*M* = 536ms, *SD* = 143ms), *t*s > 3.06, *p*s < .003. Participants also made fewer errors in the manikin task (*M* = 2.46%, *SD* = 4.22%) than in the avatar consequences task (*M* = 4.02%, *SD* = 6.40%), *t* = 2.49, *p* = .013.

**Outcome measure 1: Consumer choice.** A Kruskal-Wallis test revealed a main effect of AA Task Condition on behavioral choice scores, $\chi^2(2) = 33.69$, $p < .001$. Multiple comparisons indicated that the scores were higher for participants in the goal-relevant avatar consequences task condition (*M* = 0.63, *SD* =0.51) than for participants in the avatar consequences task (*M* = 0.47, *SD* =0.62) or manikin task (*M* = 0.24, *SD* =0.58) conditions, *p*s <.035. Replicating the

results of Experiment 1, the scores were higher for the avatar consequences task than for the manikin task, $p < .001$. Participants in all task conditions selected the positive brand more often than the negative brand, $p$s <.001, $BF_{1}$s > 10000 (manikin: $d_z = 0.42$, avatar consequences: $d_z = 0.75$, goal-relevant avatar consequences: $d_z = 1.22$).

**Outcome measure 2: IAT scores.** Spearman-Brown corrected split-half reliability of the (personalized) IAT score was $r(453) = .90$. The ANOVA revealed a main effect of IAT Block Order, $F(1,443) = 12.30$, $p$ <.001, as well as a main effect of Positive Brand, $F(1,443) = 7.30$, $p$ =.007. More importantly, we also observed a main effect of AA Task Condition, $F(2,443) = 9.02$, $p$ <.001. IAT scores were higher for participants in the goal-relevant avatar consequences task condition ($M = 0.40$, $SD = 0.37$) than for participants in the avatar consequences task ($M = 0.23$, $SD = 0.44$) or manikin task ($M = 0.28$, $SD$ =0.38) conditions, $t$s > 2.75, $p$s < .007. IAT scores were not significantly different for avatar consequences task and manikin task conditions, $t(246) = 1.12$, $p = .27$. Participants in all task conditions exhibited an implicit preference for the positive over the negative brand, $p$s < .001, $BF_{1s} > 10^5$ (manikin: $d_z = 0.72$; avatar consequences: $d_z = 0.52$, goal-relevant avatar consequences: $d_z = 1.07$).

**Outcome measure 3: Explicit rating scores.** An ANOVA on explicit rating scores revealed the main effect of AA Task Condition, $F(2,449) = 65.89$, $p < .001$. Explicit rating scores were higher for the goal-relevant avatar consequences task condition ($M = 5.60$, $SD$ =2.71) than for the avatar consequences task condition ($M = 3.49$, $SD = 3.68$) or the manikin task condition ($M = 1.75$, $SD = 2.38$), $t$s > 5.60, $p$s < .001. Replicating the results of Experiment 1, explicit rating scores were significantly higher for participants in the avatar consequences task condition than for participants in the manikin task condition, $t(246) = 4.85$, $p$ <.001. Participants in all task conditions exhibited a preference for the positive brand over the negative brand, $p$s <.001, $BF_{1}$s >

$10^5$ (manikin: $d_z$ = 0.74; avatar consequences: $d_z$ = 0.95, goal-relevant avatar consequences: $d_z$ = 2.07).

**Outcome measure 4: Approach-avoidance scores.** Spearman-Brown corrected split-half reliability of the approach-avoidance score was $r(453) = .93$. The ANOVA on approach-avoidance scores also revealed the main effect of AA Task Condition, $F(2,449) = 40.33$, $p < .001$. Scores were higher for the avatar consequences task condition ($M = 4.09$, $SD = 1.77$) compared to avatar consequences task ($M = 2.58$, $SD = 3.09$) and manikin task ($M = 1.53$, $SD = 2.61$) conditions, $t$s > 5.14, $p$s < .001. In line with Experiment 1, approach-avoidance scores were higher for participants in the avatar consequences task condition than for participants in the manikin task condition, $t(285) = 3.58$; $p < .001$. Overall, participants in all task conditions approached the positive brand more often than the negative brand, $p$s <.001, $BF_1$s > $10^5$ (manikin: $d_z$ = 0.59, avatar consequences: $d_z$ = 0.83, goal-relevant avatar consequences: $d_z$ = 2.32).

**Reactance scores.** An ANOVA on reactance scores revealed a main effect of AA Task Condition, $F(2,415) = 22.91$, $p < .001$. Participants gave higher reactance ratings in the manikin task condition ($M = 4.46$, $SD = 2.48$) than in the avatar consequences task ($M = 3.65$, $SD = 2.31$) or goal-relevant avatar consequences task condition ($M = 2.61$, $SD = 2.19$), $t$s > 2.90, $p$s < .004 (Table 1). Reactance scores were also higher for participants in the avatar consequences task than for participants in the goal-relevant avatar consequences task condition, $t(265) = 3.68$, $p < .001$. We did not observe any significant differences in correlations between reactance scores and IAT or explicit rating scores between the difference task conditions (overall correlation between reactance scores and explicit rating scores: $r(416) = -.35$, $p < .001$; reactance scores and IAT scores: $r(416) = -.19$, $p < .001$).

**Demand.** A Chi-square proportion test indicated that a significantly higher proportion of participants in the manikin task condition (9.6%) indicated that they had complied to experimental demands when providing their evaluative ratings than in the goal-relevant avatar consequences task (2.0%), $\chi^2(1) = 6.77$, $p = .009$ (Table 1). We observed no significant differences between the avatar consequences task condition and the other conditions, $\chi^2$s < 2.39, *ps* >.12. Explicit rating scores and IAT scores did not differ significantly between demand compliant and non-demand compliant participants in any of the conditions, *F*s < 0.96, *p*s > .33. Excluding the data of demand compliant participants did not change the statistical significance of any of the reported effects.

**Contingency awareness.** A Chi-square proportion test indicated that a significantly higher proportion of participants in the manikin task (79.6%) and goal-relevant avatar consequences (90.1%) conditions indicated correct contingencies than in the avatar consequences task (61.0%), $\chi^2$s > 11.81, *ps* < .001.

**Results of Experiment 3**

**AA Task.** We examined participants' mean latencies and errors in the AA task. Between-subjects *t*-tests revealed that participants were faster in the manikin task (*M* = 555ms, *SD* = 183ms) than in the control manikin task (*M* = 603ms, *SD* = 171ms) or the goal-relevant avatar consequences task (*M* = 615ms, *SD* = 156ms), *t*s > 2.76, *p*s < .007. Participants did not make fewer errors in the manikin task (*M* = 4.18%, *SD* = 5.21%) than in the control manikin task (*M* = 3.87%, *SD* = 5.86%), *t* = -0.45, *p* = .013.

**Pre-task questions.** Participants in the three task conditions did not provide significantly different ratings for Healthy Eating Intention, *t*s < 1.30, *p*s > .19, Hunger, *t*s < 0.38, *p*s >.70,

Healthy Eating Behavior, *t*s < 1.61, *p*s >.10, or Healthy Eating Behavior Difficulty, *t*s < 0.93, *p*s >.35.

**Outcome measure 1: Consumer choice.** We computed two behavioral choice scores by subtracting the number of times participants chose a coupon for the healthy item over the unhealthy item in the behavioral choice task for trade-off choices and for control choices. Behavioral choice scores were subjected to an ANOVA with AA Task Condition and Choice (Trade-off, Control) as between-subjects factor and Health Behavior, Health Intention, Health Behavior Difficulty and Hunger as covariates. We observed main effects of Health Behavior, Health Intention, and Hunger, *F*s > 6.40, *p*s < .012. We also observed a main effect of Choice, $F(1,386) = 76.20$, $p < .001$. Participants chose the coupon for the healthy food less for trade-off choices, $BF_1 > 10^5$, $d = 0.72$. There was no interaction of Choice and AA Task Condition, $F(2,386) = 2.14$, $p = .12$. Importantly, however, we did observe a main effect of AA Task Condition, $F(2,382) = 11.94$, $p < .001$. Contrasts revealed that behavioral choice scores were higher for the goal-relevant avatar consequences task condition ($M = 2.25$, $SD = 0.83$) than for the control task ($M = 1.78$, $SD = 0.80$) or manikin task ($M = 1.97$, $SD = 0.83$) conditions, *t*s > 2.70, *p*s < .008, $BF_1$s > 4.24, *d*s > 0.33. Behavioral choice scores were higher for the manikin task condition than for the control task condition, but this effect was only marginally significant, $t(261) = 1.84$, $p = .066$, $BF_0 = 1.48$, $d = 0.23$. Participants in all task conditions selected the healthy foods more often than the unhealthy foods, *p*s < .001, $BF_1$s > 194.19 (control: $d_z = 0.36$, manikin: $d_z = 0.57$, goal-relevant avatar consequences: $d_z = 0.90$).

**Outcome measure 2: IAT scores.** Spearman-Brown corrected split-half reliability of the personalized IAT score was $r(387) = .97$. The ANOVA that included Health Behavior, Health Intention, Health Behavior Difficulty and Hunger as covariates revealed only a main effect of AA

Task Condition, $F(2,379) = 3.83$, $p = .023$. Contrasts revealed that IAT scores were higher for the goal-relevant avatar consequences task condition ($M = 0.90$, $SD = 0.35$) than for the control task ($M = 0.78$, $SD = 0.37$) and manikin task ($M = 0.80$, $SD = 0.34$) conditions, $t$s > 2.16, $p$s < .032, $BF_1$s > 1.89, $d$s > 0.29. IAT scores did not differ significantly between manikin task and control task conditions, $t(261) = 0.41$, $p = .68$, $BF_0 = 6.80$, $d = 0.05$. Participants in all task conditions exhibited an implicit preference for healthy foods over unhealthy foods, $p$s < .001, $BF_1$s > $10^5$ (control: $d_z = 2.11$, manikin: $d_z = 2.38$, goal-relevant avatar consequences: $d_z = 2.58$).

**Outcome measure 3: Explicit rating scores.** An ANOVA on explicit rating scores revealed main effects of Health Intention, Health Behavior Difficulty, and Hunger, $F$s > 5.84, $p$s < .016, as well as the main effect of AA Task Condition, $F(2,382) = 3.08$, $p = .047$. Contrasts revealed that explicit rating scores were higher for the goal-relevant avatar consequences task condition ($M = 3.27$, $SD = 2.71$) compared to the control task condition ($M = 2.57$, $SD = 2.63$), $t(254) = 2.45$, $p = .015$, $BF_1 = 1.71$, $d = 0.29$, but not compared to the manikin task condition ($M = 2.82$, $SD = 2.53$), $t(257) = 1.57$, $p = .12$, $BF_0 = 1.38$, $d = 0.23$. The scores for the manikin task condition were not significantly different from the scores for the control task, $t(261) = 0.89$, $p = .38$, $BF_0 = 6.56$, $d = 0.06$. Participants in all task conditions indicated that they preferred healthy food over unhealthy food, $p$s <.001, $BF_1$s > $10^5$ (control: $d_z = 0.98$, manikin: $d_z = 1.08$, goal-relevant avatar consequences: $d_z = 1.23$).

**Outcome measure 4: Approach-avoidance scores.** Spearman-Brown corrected split-half reliability of the approach-avoidance score was $r(387) = .91$. The ANOVA on approach-avoidance scores revealed main effects of Health Behavior and Health Intention, $F$s > 6.19, $p$s < .014, as well as the main effect of AA Task Condition, $F(2,382) = 5.25$, $p = .006$. Scores were higher for participants in the goal-relevant avatar consequences task condition ($M = 2.84$, $SD$

=2.52) and for participants in the manikin task condition ($M$ = 2.56, $SD$ = 2.68) compared to participants in the control task condition ($M$ = 1.86, $SD$ = 2.82), $t$s > 2.24, $p$s < .026, $BF_1$s > 1.78, $d$s > 0.21. Scores were not significantly different between manikin task and avatar consequences task conditions, $t(257) = 0.91$, $p$ =.36, $BF_0 = 3.09$, $d = 0.17$. Participants in all task conditions approached healthy foods more than unhealthy foods, $p$s <.001, $BF_1$s > $10^5$ (control: $d_z = 0.67$, manikin: $d_z = 0.92$, goal-relevant avatar consequences: $d_z = 1.16$).

**Reactance scores.** An ANOVA on reactance scores revealed a marginally significant main effect of Health Intention, $F(1,299) = 3.47$, $p = .064$, and a main effect of AA Task Condition, $F(2,299) = 6.63$, $p = .002$. Participants gave higher reactance ratings in the control task condition ($M$ = 4.87, $SD$ = 2.46) than in the manikin task ($M$ = 4.11, $SD$ = 2.33) or goal-relevant avatar consequences task ($M$ = 3.65, $SD$ = 2.35) conditions, $t$s > 2.26, $p$s < .025, $BF_1$s > 1.37, $d$s > 0.30 (Table 1). The difference in ratings was not significant for the goal-relevant avatar consequences task and manikin task conditions, $t(209) = 1.42$, $p = .16$, $BF_0 = 2.13$, $d = 0.22$. Reactance scores correlated negatively with explicit rating scores, $r(304) = -.22$, $p < .001$, but not with IAT scores, $r(304) = -.02$, $p = .79$. Correlation coefficients for reactance and explicit rating scores differed significantly but only between manikin task condition ($r[108] = -.43$) and goal-relevant avatar consequences condition ($r[99] = -.09$), $p = .044$. Correlation coefficients for reactance and IAT scores differed significantly between manikin task condition ($r[108] = -.18$) and both control ($r[93] = .11$) and goal-relevant avatar consequences task ($r[99] = .10$) conditions, $p$s < .043.

**Demand.** A Chi-square proportion test did not reveal significant differences in the proportion of participants who were demand compliant in the manikin task (3.3%), control task (0.8%), or goal-relevant avatar consequences task (2.6%), $\chi^2$s < 1.88, $p$s >.39 (Table 1). Explicit

rating scores and IAT scores did not differ significantly between demand compliant and non-demand compliant participants in any of the conditions, *F*s < 0.44, *p*s > .51. Excluding the data of demand compliant participants did not change the statistical significance of any of the reported effects.

**Contingency awareness.** A Chi-square proportion test indicated no significant differences in the proportion of participants who indicated correct contingencies in the manikin task (69.6%) and goal-relevant avatar consequences (76.9%), $\chi^2(1) = 1.44$, $p = .23$.

**Phase 2 questions.** A multivariate ANOVA on all phase 2 question ratings revealed main effects of Health Behavior, Health Behavior Difficulty, and Health Intention, *F*s > 11.07, *p*s < .001, as well as a main effect of AA Task Condition, $F(10,301) = 4.46$, $p < .001$.

An ANOVA on healthy eating behavior ratings revealed main effects of Health Behavior and Health Intention, *F*s > 11.07, *p*s < .001, but no main effect of AA Task Condition, $F(2,301) = 0.95$, $p = .39$. Prompted by the suggestion of a reviewer, we deviated from our data analysis plan by performing an additional ANOVA on healthy eating behavior ratings collected before the AA task (Phase 1) and after the one day delay (Phase 2). This analysis revealed a significant main effect of AA Task Condition, $F(2,302) = 10.08$, $p < .001$, and a main effect of Phase, $F(1,304) = 33.21$, $p < .001$, but their interaction was not significant, $F(2,304) = 0.48$, $p = .62$. Further exploration of the AA Task Condition effect revealed that, overall, healthy eating behavior ratings were higher for the goal-relevant avatar consequences task condition ($M = 5.29$, $SD = 1.70$) than for the manikin task condition ($M = 5.98$, $SD = 1.72$).

An ANOVA on unhealthy eating behavior ratings revealed main effects of Health Behavior and Health Behavior Difficulty, *F*s > 3.63, *p*s < .028, as well as the main effect of AA Task Condition, $F(2,301) = 3.63$, $p = .028$. Contrasts revealed that scores were lower for the goal-

relevant avatar consequences task condition ($M$ = 4.13, $SD$ = 1.79) than for control task ($M$ = 4.63, $SD$ = 1.78) and manikin task ($M$ = 4.68, $SD$ = 1.89) conditions, $t$s > 2.23, $p$s < .027, $BF_1$s > 5.08, $d$s > 0.31. The scores did not differ for the manikin task and control task conditions, $t(205) = 0.84$, $p = .40$, $BF_0 = 4.73$, $d = 0.12$. We did not perform an additional ANOVA on unhealthy eating ratings in Phase 1 and Phase 2 because these ratings were not collected in Phase 1

An ANOVA on healthy eating behavior intention ratings revealed main effects of Health Behavior and Health Intention, $F$s > 28.72, $p$s < .001, as well as the main effect of AA Task Condition, $F(2,308) = 3.31$, $p = .038$. Contrasts revealed that scores were higher for the goal-relevant avatar consequences task condition ($M$ = 6.41, $SD$ = 2.06) compared to the manikin task condition ($M$ = 6.02, $SD$ = 2.29), $t(206) = 2.93$, $p$ =.004, $BF_1 = 8.03$, $d = 0.41$ and the control task condition ($M$ = 5.78, $SD$ = 2.28), $t(206) = 2.55$, $p$ =.002, $BF_1 = 16.68$, $d = 0.44$. The scores did not differ between manikin task and control task conditions, $t(205) = 0.21$, $p$ =.83, $BF_0 = 6.47$, $d$ = 0.03, $BF_0 = 6.47$, $d = 0.03$. An additional ANOVA on healthy eating behavior intention ratings at Phase 1 and Phase 2 revealed a significant main effect of AA Task, $F(2,302) = 9.42$, $p < .001$, a main effect of Phase, $F(1,304) = 31.99$, $p < .001$, and a significant Phase x AA Task interaction, $F(2,304) = 3.17$, $p$ =.043. The interaction revealed that the main effect of AA Task, indicating a higher rating for the goal-relevant avatar consequences task condition than for the manikin and control task conditions, was only present at Phase 2 ($p$s < .004).

An ANOVA on ratings referring to participants' difficulty to stop eating unhealthily revealed only a main effect of Health Behavior Difficulty, $F(1,301) = 130.78$, $p$s < .001, but not a main effect of AA Task Condition, $F(2,301) = 0.25$, $p = .78$. An additional ANOVA on healthy eating difficulty ratings at Phase 1 and Phase 2 revealed a significant main effect of Phase, $F(1,304) = 82.37$, $p < .001$, but no main or interaction effects of AA Task, $F$s < 1.33, $p$s > .25.

An ANOVA on healthy eating goal ratings revealed main effects of Health Behavior and Health Intention, *F*s > 26.94, *p*s < .001, and a marginally significant main effect of AA Task Condition, $F(2,301) = 2.76$, $p = .065$. Contrasts revealed that the scores were higher for the goal-relevant avatar consequences task condition ($M = 6.96$, $SD = 1.56$) compared to the control task condition ($M = 6.54$, $SD = 1.96$), $t(197) = 2.29$, $p = .023$, $BF_1 = 3.28$, $d = 0.36$, but not compared to the manikin task condition ($M = 6.67$, $SD = 1.97$), $t(205) = 1.60$, $p = .11$, $BF_1 = 34.56$, $d = 0.48$. The scores did not differ between manikin task and control task conditions, $t(205) = 0.70$, $p = .48$, $BF_0 = 5.05$, $d = 0.10$. An additional ANOVA on healthy eating goal ratings in Phase 1 and Phase 2 revealed a significant main effect of AA Task Condition, $F(2,302) = 8.49$, $p < .001$ and a marginally significant Phase x AA Task Condition interaction, $F(2,304) = 2.33$, $p = .099$. The interaction revealed that the main effect of AA Task Condition, indicating a higher rating for the goal-relevant avatar consequences task condition than for the manikin and control task conditions, was only present in Phase 2 (*p*s < .012).

**Results of Experiment 4**

**AA Task.** We examined participants' mean latencies and errors in the AA task. Between-subjects *t*-tests did not reveal significant differences in latencies in the control manikin task ($M =$ 533ms, $SD =$ 124ms) and the goal-relevant avatar consequences task ($M =$ 557ms, $SD =$ 42ms), $t(182) = 1.73$, $p = .085$. Participants' mean errors in the control manikin task was 4.68% ($SD =$ 6.17%).

**Pre-task questions.** Participants in the two task conditions did not provide significantly different ratings for Healthy Eating Intention, $t(182) = 0.39$, $p = .70$, Hunger, $t(182) = 0.57$, $p =$ .57, Healthy Eating Behavior, $t(182) = 0.00$, $p > .99$, or Healthy Eating Behavior Difficulty, $t(182) = 1.09$, $p = .28$.

**Outcome measure 1: Consumer choice.** We computed two behavioral choice scores by subtracting the number of times participants chose the healthy item over the unhealthy item in the behavioral choice task for trade-off choices and for control choices. Behavioral choice scores were subjected to an ANOVA with AA Task Condition and Choice (Trade-off, Control) as between-subjects factor and Health Behavior, Health Intention and Hunger as covariates. We observed main effects of Health Behavior, Health Intention, and Hunger, $F$s > 5.30, $p$s < .023. We also observed a main effect of Choice, $F(1,182) = 16.32$, $p$ <.001. Participants chose the healthy food less for trade-off choices, $BF_1 = 188.91$, $dz = 0.36$. There was no interaction of Choice and AA Task Condition, $F(1,182) = 0.07$, $p = .79$. Importantly, however, we did observe a main effect of AA Task Condition, $F(1,179) = 8.91$, $p = .003$. A planned contrast based on a one-sided $t$-test revealed that behavioral choice scores were higher for the goal-relevant avatar consequences task condition ($M = 1.88$, $SD = 0.80$) than for the control task condition ($M = 1.59$, $SD = 0.68$), $t(182) = 2.59$, $p = .005$, $BF_1 = 7.77$, $d = 0.38$. Participants selected the healthy foods more often than the unhealthy foods in the avatar consequences task condition, $t(91) = 4.52$, $p$ < .001, $BF_1 = 883.87$, $d = 0.47$, but not the control task condition, $t(91) = 1.31$, $p$ =.20, $BF_0 = 3.82$, $d = 0.14$.

**Outcome measure 2: IAT scores.** Spearman-Brown corrected split-half reliability of the personalized IAT score was $r(182) = .98$. The ANOVA revealed main effects of Health Intention and IAT Block Order, $F$s > 9.40, $p$s < .003, and, crucially, also the main effect of AA Task Condition, $F(1,180) = 4.96$, $p = .027$. Contrasts revealed that IAT scores were higher for the goal-relevant avatar consequences task condition ($M = 0.87$, $SD = 0.25$) than for the control task condition ($M = 0.77$, $SD = 0.37$), $t(182) = 1.97$, $p$ =.025, $BF_1 = 2.76$, $d = 0.29$. Participants in both

task conditions exhibited an implicit preference for healthy foods over unhealthy foods, *p*s < .001, $BF_{1s} > 10^5$ (control: $d_z = 2.12$, goal-relevant avatar consequences: $d_z = 3.50$).

**Outcome measure 3: Explicit rating scores.** The ANOVA on explicit ratings revealed main effects of Health Intention and Health Behavior, *F*s > 8.09, *p*s < .005, but no main effect of AA Task Condition, $F(1,180) = 0.06$, $p = .81$. Contrasts did not reveal higher explicit rating scores for the goal-relevant avatar consequences task condition ($M = 4.20$, $SD = 2.09$) compared to the control task condition ($M = 4.17$, $SD = 2.15$), $t(182) = 0.07$, $p = .47$, $BF_0 = 2.80$, $d = 0.01$. Participants in all task conditions indicated that they preferred healthy food over unhealthy food, *p*s <.001, $BF_1$s > $10^5$ (control: $d_z = 1.95$, goal-relevant avatar consequences: $d_z = 1.99$).

**Outcome measure 4: snack eating behavior in grams.** The ANOVA on snack eating revealed main effects of Health Behavior and Hunger, *F*s > 5.30, *p*s < .023, as well as the main effect of AA Task Condition, $F(1,180) = 5.67$, $p = .018$. Participants consumed less grams of food in the goal-relevant avatar consequences task condition ($M = 36.49$, $SD = 25.22$) than in the control task condition ($M = 46.72$, $SD = 33.53$), $t(182) = 2.34$, $p = .010$, $BF_1 = 5.33$, $d = 0.34$.

**Outcome measure 5: snack eating behavior in calories.** The ANOVA on snack eating revealed main effects of Health Behavior and Hunger, *F*s > 4.29, *p*s < .040, as well as the main effect of AA Task Condition, $F(1,180) = 4.81$, $p = .030$. Participants consumed less calories in the goal-relevant avatar consequences task condition ($M = 1.56$ $SD = 1.08$) than in the control task condition ($M = 1.94$, $SD = 1.38$), $t(182) = 2.18$, $p = .015$, $BF_1 = 4.16$, $d = 0.32$.

**Outcome measure 6: Temptation in snack eating task.** The ANOVA on temptation ratings revealed main effects of Health Behavior, and Hunger, *F*s > 5.58, *p*s < .020, but no main effect of AA Task Condition, $F(1,179) = 0.09$, $p = .77$. Contrasts did not reveal lower temptation

ratings for the goal-relevant avatar consequences task condition ($M$ = 3.94, $SD$ = 1.85) compared to the control task condition ($M$ = 4.01, $SD$ = 1.86), $t(182) = 0.40$, $p = .35$, $BF_0 = 2.50$, $d = 0.06$.

**Reactance scores.** An ANOVA on reactance scores revealed only a main effect of AA Task Condition, $F(1,119) = 6.27$, $p = .014$. Participants gave higher reactance ratings in the control task condition ($M$ = 3.79, $SD$ = 2.30) than in the goal-relevant avatar consequences task condition ($M$ = 2.82, $SD$ = 1.95), $t(119) = 2.50$, $p = .007$, $BF_1 = 7.44$, $d = 0.46$ (Table 1). Reactance scores did not correlate significantly with explicit rating scores or IAT scores, $p$s > .23.

**Demand.** There were no participants who were demand compliant in any of the two task conditions (Table 1).

**Contingency awareness.** 86.96 % of participants indicated correct contingencies in the goal-relevant avatar consequences.

**Task engagement ratings.** A between-subjects $t$-test revealed that participants had lower self-reported task engagement in the control task condition ($M$ = 4.23, $SD$ = 0.57) than in the goal-relevant avatar consequences task condition ($M$ = 4.56, $SD$ = 0.45), $t(182) = 4.38$, $p < .001$, $BF_1 = 896.15$, $d = 0.65$. Task engagement scores were significantly higher than the rating of 4 (indicating very high task engagement) in both conditions, $t$s > 3.79, $p$s < .001. Task engagements scores did not correlate significantly with any of the outcome measures, with the exception of IAT scores, $r(182) = 0.23$, $p = .002$, but the effect of AA Task on IAT scores was not moderated by task engagement, $F(1,178) = 0.49$, $p = .49$.

## References

Blechert, J., Meule, A., Busch, N.A., & Ohla, K. (2014). Food-pics: an image database for experimental research on eating and appetite. *Frontiers in Psychology,* 5:617. doi: 10.3389/fpsyg.2014.00617.

Cacioppo, J. T., Priester, J. R., & Berntson, G. G. (1993). Rudimentary determinants of attitudes II: Arm flexion and extension have differential effects on attitudes. *Journal of Personality and Social Psychology, 65,* 5–17. doi:10.1037/0022-3514.65.1.5

Charbonnier, L., van Meer, F., van der Laan, L.N., Viergever, M.A., Smeets, P.A. (2016). Standardized food images: A photographing protocol and image database. *Appetite,* 96:166–173.

Epstein L.H., Dearing, K. K. & Roba, L. G. (2010). A questionnaire approach to measuring the relative reinforcing efficacy of snack foods. Eating Behavior, 11, 67–73. doi:10.1016/j.eatbeh.2009.09.006.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175-191.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85,* 197–216. doi: 10.1037/0022-3514.85.2.197

Han, H. A., Olson, M. A., & Fazio, R. H. (2006). The influence of experimentally-created extra-personal associations on the Implicit Association Test. *Journal of Experimental Social Psychology, 42,* 259-272.

Haynes, A., Kemps, E., & Moffitt, R. (2016). Too Depleted to Try? Testing the Process Model of Ego Depletion in the Context of Unhealthy Snack Consumption. *Applied Psychology: Health and Well-Being, 8(3*), 386–404

Haynes, A., Kemps, E., & Moffitt, R. (2015). Inhibitory self-control moderates the effect of changed implicit food evaluations on snack food consumption. *Appetite, 90,* 114–122

Helton, W. S., & Näswall, K. (2015). Short Stress State Questionnaire: Factor structure and state change assessment. *European Journal of Psychological Assessment, 31*(1), 20-30.doi:10.1027/1015-5759/a000200

Hughes, S. Van Dessel, P. & Smith, C. T. (2017). A comparative analysis of different evaluative learning procedures. Manuscript in preparation.

Kakosche, N., Kemps, E., & Tiggeman, M. (2018). What is the appropriate control condition for approach bias modification? A response to commentary by Becker et al. (2017). *Addictive behaviors, 77,* 295–296

Krieglmeyer, R., & Deutsch, R. (2010). Comparing measures of approach-avoidance behaviour: The manikin task vs. two versions of the joystick task. *Cognition & Emotion, 24* , 810-828.

Krieglmeyer, R., Deutsch, R., De Houwer, J., & De Raedt, R. (2010). Being Moved: Valence Activates Approach-Avoidance Behavior Independently of Evaluation and Approach-Avoidance Intentions. *Psychological Science*, *21*(4), 607–613. https://doi.org/10.1177/0956797610365131

Krieglmeyer, R., De Houwer, J., & Deutsch, R. (2013). On the nature of automatically triggered approach–avoidance behavior. *Emotion Review, 5,* 280-284.

Mai, R., Hoffmann, S., Hoppert, K., Schwarz, P., & Rohm, H. (2015). The spirit is willing, but the flesh is weak: The moderating effect of implicit associations on healthy eating behaviors. *Food Quality and Preference, 39,* 62–72. doi:10.1016/j.foodqual.2014.06.014

Markman, A. B., & Brendl, C. M. (2005). Constraining theories of embodied cognition. *Psychological Science, 16,* 6-10. doi: 10.1111/j.0956-7976.2005.00772.x

Morwitz, Vicki G. (2001), "Methods for Forecasting from Intentions Data," in Principles of Forecasting: A Handbook for Researchers and Practitioners, J. Scott Armstrong, ed. Norwell, MA: Kluwer Academic Publishers, 34–56.

Neumann, R., & Strack, F. (2000). Approach and avoidance: the influence of proprioceptive and exteroceptive cues on encoding of affective information. *Journal of Personality and Social Psychology, 79,* 39–48. doi: 10.1037/0022-3514.79.1.39

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16,* 225–237. doi: 10.3758/PBR.16.2.225

Salmon, S. J., Fennis, B. M., de Ridder, D. T. D., Adriaanse, M. A., & de Vet, E. (2014). Health on impulse: When low self-control promotes healthy food choices. *Health Psychology, 33,* 103–109. http://dx.doi.org/10.1037/a0031785.

Sharpe, D., & Whelton, W. J. (2016). Frightened by an Old Scarecrow: The remarkable resilience of demand characteristics. *Review of General Psychology*.

Smith, C. T., De Houwer, J., & Nosek, B. (2013). Consider the source: persuasion of implicit evaluations is moderated by source credibility. *Personality and Social Psychology Bulletin,* 39, 193-205. doi:10.1177/0146167212472374

Van Dessel, P., De Houwer, J., & Gast, A. (2016). Approach-Avoidance Training effects are Moderated by Awareness of Stimulus-Action Contingencies. *Personality and Social Psychology Bulletin, 42*, 81-93. doi:10.1177/0146167215615335

Van Dessel, P., De Houwer, J., Gast, A. & Spruyt, A. (2017). Approach-Avoidance Training effects: The moderating role of stimulus valence and action task. Manuscript in preparation.

Van Dessel, P., Eder, A. B., & Hughes, S. (in press). Mechanisms Underlying Approach-Avoidance Training Effects on Stimulus Evaluation. *Journal of Experimental Psychology: Learning, Memory, & Cognition.*

Van Dessel, P., Gawronski, B., Smith, C. T., & De Houwer, J. (2017). Mechanisms underlying approach-avoidance instruction effects on implicit evaluation: Results of a preregistered adversarial collaboration. *Journal of Experimental Social Psychology, 69*, 23-32. doi:10.1016/j.jesp.2016.10.004

Van Dessel, P., Hughes, S., & De Houwer, J. (2018). How Do Actions Influence Attitudes? An Inferential Account of the Impact of Action Performance on Stimulus Evaluation. Manuscript invited for revision at Personality and Social Psychology Review. Prepint available at: https://osf.io/kb3wq/

Woud, M. L., Maas, J., Becker, E. S., & Rinck, M. (2013). Make the manikin move: Symbolic approach–avoidance responses affect implicit and explicit face evaluations. *Journal of Cognitive Psychology*, *25*, 738–744. doi: 10.1080/20445911.2013.817413

Zanon, R., De Houwer, J., Gast, A., & Smith, C. T. (2014). When does relational information influence evaluative conditioning? *Quarterly Journal of Experimental Psychology, 67,* 2105-2122. doi:10.1080/17470218.2014.907324

Zhou, H., & Fishbach, A. (2016). The Pitfall of Experimenting on the Web: How Unattended Selective Attrition Leads to Surprising (Yet False) Research Conclusions. *Journal of Personality and Social Psychology, 11,* 493-504. doi: 10.1037/ pspa0000056

## Tables

Table 1. *Mean reactance and demand compliance scores in Experiments 1, 2, and 3.*

| | Manikin task | Manikin control task | Avatar task | Avatar consequences task | Goal-relevant avatar consequences task |
|---|---|---|---|---|---|
| **Reactance** | **Mean (SD)** | **Mean (SD)** | **Mean (SD)** | **Mean (SD)** | **Mean (SD)** |
| Experiment 1 | 4.23 (2.65) | / | 3.86 (2.44) | 3.41 (2.34) | |
| Experiment 2 | 4.46 (2.48) | / | / | 3.65 (2.31) | 2.61 (2.19) |
| Experiment 3 | 4.11 (2.33) | 4.87 (2.46) | / | / | 3.65 (2.35) |
| Experiment 4 | / | 3.79 (2.30) | / | / | 2.82 (1.92) |
| **Demand compliance** | **Proportion** | **Proportion** | **Proportion** | **Proportion** | **Proportion** |
| Experiment 1 | 9.2% | / | 8.8% | 2.9% | / |
| Experiment 2 | 9.6% | / | / | 6.2% | 1.8% |
| Experiment 3 | 3.3% | 0.8% | / | / | 2.6% |
| Experiment 4 | / | 0.0% | / | / | 0.0% |