

Progress report – Bial Foundation

Are free will and moral responsibility real or illusory? On the causal role of consciousness in decision-making, a combined EEG and intracranial study

PI: Uri Maoz

Keywords: Decision-Making, Volition, Free Will, Moral Responsibility, and Role of Consciousness in Decision-Making

Human beings generally experience their decisions and actions as up to them in a meaningful way. They decide, for example, whether to donate money to charity or use it to go on vacation, and they are therefore typically held morally responsible for the consequences. But some relatively recent neuroscientific studies claimed to challenge that notion, showing that there may be predictive information in the brain about upcoming decisions before the person reported having consciously decided (Fried, Mukamel, & Kreiman, 2011; Haggard & Eimer, 1999; Libet, Gleason, Wright, & Pearl, 1983; Soon, Brass, Heinze, & Haynes, 2008; Soon, He, Bode, & Haynes, 2013). Therefore, some have claimed, it is unconscious brain activity that initiates action, and the conscious decision may be ineffective, or irrelevant to the production of action. If this interpretation is true, our introspective experience of free will may be no more than an alluring illusion (Harris, 2012; Libet, 1985; Mele, 2006, 2009; Roskies, 2010; Sinnott-Armstrong & Nadel, 2011; Wegner, 2002).

However, these experiments have also come under conceptual and empirical criticism. Importantly, the type of decisions that were studied in them—e.g., arbitrarily raising the left or right hand at a time of the subject's choice for no reason or purpose and with no consequence to the subject—are highly artificial. In everyday situations, outside the lab, humans generally act for reasons, and their decisions tend to have consequences. In that vein, it would be ludicrous to take someone to court for having raised her left hand rather than her right hand for no reason or purpose. But if she sends someone to the gallows by lifting her left hand or sets him free by lifting her right, there are moral implications to her decision (Maoz & Yaffe, 2013).

We previously investigated monkeys deciding between smaller, immediate rewards and larger, delayed ones. We found single-unit activity in the dorsolateral prefrontal cortex and striatum

that predicted the choice of the animals before they were even shown the visual cues that represented the decision alternatives. This early activity was more predictive the more the values of the decision alternatives for the monkeys were similar. We therefore interpreted this as bias activity that integrates together with the activity resulting from the values of the decision alternatives to form the final decision. When the values of the decision alternatives were divergent, the bias made little difference in the decision outcome; when the values were similar, the bias had a greater role in the decision (Maoz et al., 2013). Accordingly, we hypothesized that the bias would have a greater effect on the decision outcome in deliberate decisions—reasoned, meaningful and purposeful decisions that bear consequences for the subject—than on random decisions—unreasoned, meaningless, purposeless and bearing no consequences. This, together with other, more-recent work, suggests the distinct neural mechanisms underlie deliberate and random decisions.

Therefore, it seems that the specific, previous claims that were made, stating that consciousness has no causal role in decision-making, hold no merit. But even if we ignore those specific claims, the role of consciousness in decision-making, if any, remains unclear (Newell & Shanks, 2014). A critical step toward deciphering this role is a rigorous and comprehensive investigation of the temporal relation between awareness of decisions and the neural signals predictive of these decisions. This was the key point of our proposal and is now the focus of our work.

We are currently at various stages of different studies probing the causal role of consciousness in decision-making. First and chief among them is the study that was proposed in the application to the Bial Foundation. That study included 3 tasks. In Task 1 subjects were to play a matching-pennies game against the computer, where they would need to press the left or right button with their left or right hand, respectively, immediately at the go signal. Importantly, 20% of the trials would be catch trials, where the go signal would arrive unexpectedly before the end of the countdown. In those trials, subjects would indicate their degree of confidence in having already decided at the time of the unexpected go signal. The idea would be to record EEG during this process and to further compare deliberate decisions (as in the game above) to arbitrary ones (where the subjects would be explicitly informed that they are playing against a random-number generator with no patterns in its selections).

We programmed this paradigm and gathered behavioral data (N = 12 subjects). However, we found several problems with the paradigm. First, the catch trials, though rare, severely interrupted the flow of the game and confused subjects. Second, in post-experiment briefings, subjects reported that they did not really deliberate during what we designated deliberate trials. Instead, they often experienced raising one of their hands randomly or at least without any clear reasons. And they reported that this behavior was exacerbated by the catch trials. Hence it would have been difficult to compare deliberate and random decisions in this experiment. We further tried to run subjects who were informed that they were playing against a random-number generator (N=5). But, despite our assurances, all subjects reported in post-experiment briefings that they nevertheless sought patterns in their opponent's behavior.

These combined results convinced us that the matching-pennies game is not suitable for studying the role of consciousness in decision-making in the manner we proposed to the Bial Foundation. We understood that such a competitive environment does not lend itself to the addition of catch trials or to comparing deliberate and arbitrary decisions. Instead, we decided to develop two new experimental paradigms to study the same research question. The first one emphasizes important decisions. Subjects were first presented with a list of non-profit organizations (NPOs), each with the cause they support, and they rated how much they would like to donate \$1000 (one thousand dollars) to each organization. Then, in each trial of the main experiment, the subjects saw a pair of NPOs that were either divergently rated (easy decisions) or similarly rated (hard decisions). Further, there were two types of decisions, in a blocked design. There were deliberate decisions, where one choice from one subject was randomly selected to result in a \$1000 donation to the cause the subject selected in that trial. So, the choices of the subjects determined the NPO that would be awarded the \$1000. And there were arbitrary decisions, where both NPOs in the selected trials received \$500. So, it did not matter which NPO the subject would select. In both blocks, memory catch-trials would pop up from time to time to test the subject about which of 4 NPOs appeared in the previous trial. This is to better equate memory, attention, and other cognitive resources between the two types of decisions.

Behaviorally, our manipulation worked, in the sense that deliberate decisions had longer reaction times and were more consistent with the ratings in the first part of the experiment than arbitrary decisions. For deliberate decisions, easy decisions were further faster and more consistent than hard ones (Figure 1). We therefore concluded that our manipulation worked.

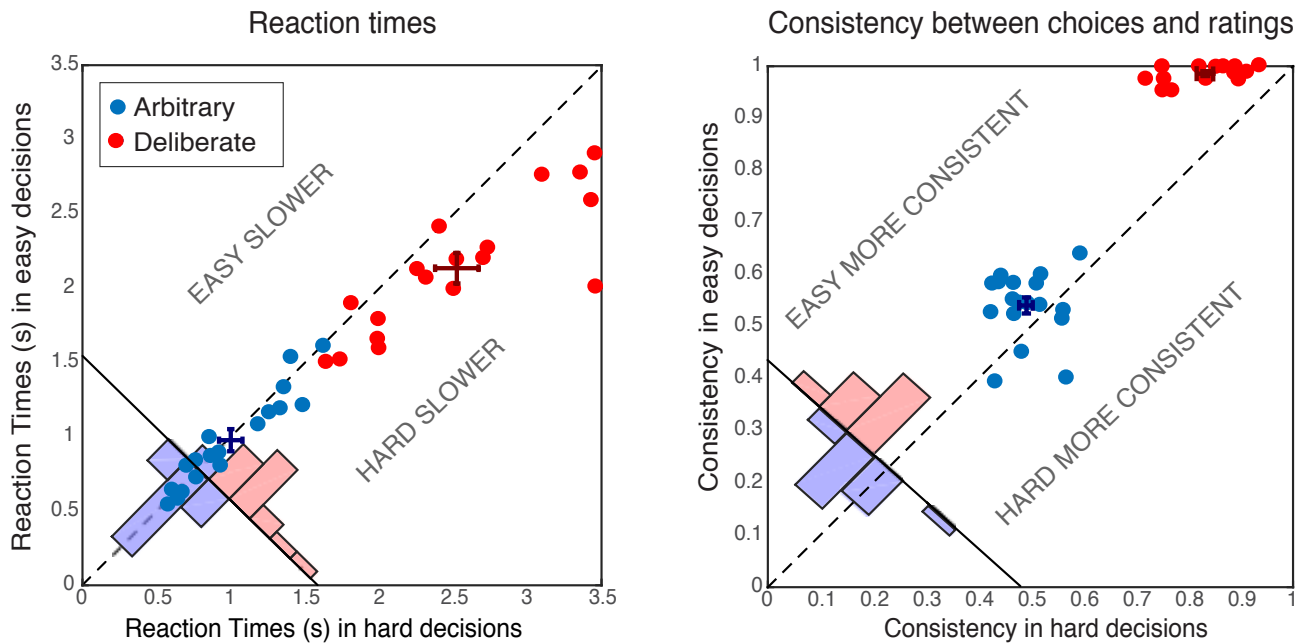


Figure 1: Behavioral results. Response Times (RTs; left) and Consistency Grades (CG; right) in arbitrary (blue) and deliberate (red) decisions. Each dot represents the average RT/CG for easy and hard decisions for an individual subject (hard decisions: x-coordinate; easy decisions: y-coordinate). Group means and SEs are represented in dark red and dark blue crosses. The histograms at the bottom-left corner of each plot sum the number of dots with respect to the solid diagonal line. The dashed diagonal line represents equal RT/CG for easy and hard decisions; data points below that diagonal indicate longer RTs or higher CGs for hard decisions. In both measures, arbitrary decisions are more centered around the diagonal than deliberate decisions, showing no or substantially reduced differences between easy and hard decisions.

The most interesting finding was regarding the readiness potential (RP). The RP is a slow negative deflection that precedes voluntary action in EEG over motor areas. In the original Libet experiments, Libet and colleagues relied on the RP to demonstrate that there may be information in the brain about upcoming action (the RP) before subjects become aware of having decided (Libet, 1985; Libet et al., 1983). However, while we replicated the finding of the RP before arbitrary decisions, we found no RP before deliberate decisions. The RP was absent both across all subjects on average (Figure 2A) and when examining single subjects (Figure 2B). These results have now been written up and submitted for publication. A pre-print of the paper can be found on the bioRxiv repository (<https://www.biorxiv.org/content/early/2018/01/04/097626>).

This NPOs paradigm is the version of Task 1 from the Bial proposal that we ended up using. For the reasons given above, we think that this version of the paradigm is best suited to answer the scientific questions we posed in the proposal. We think that the Bial Foundation would be very excited to learn about these ground-breaking results that shed new light on the role of consciousness in decision-making. These results will feature in a talk that Uri Maoz will give at

the 22nd gathering of the Association for the Scientific Study of Consciousness (ASSC) in Krakow, Poland at the end of June, 2018.

Readiness Potentials

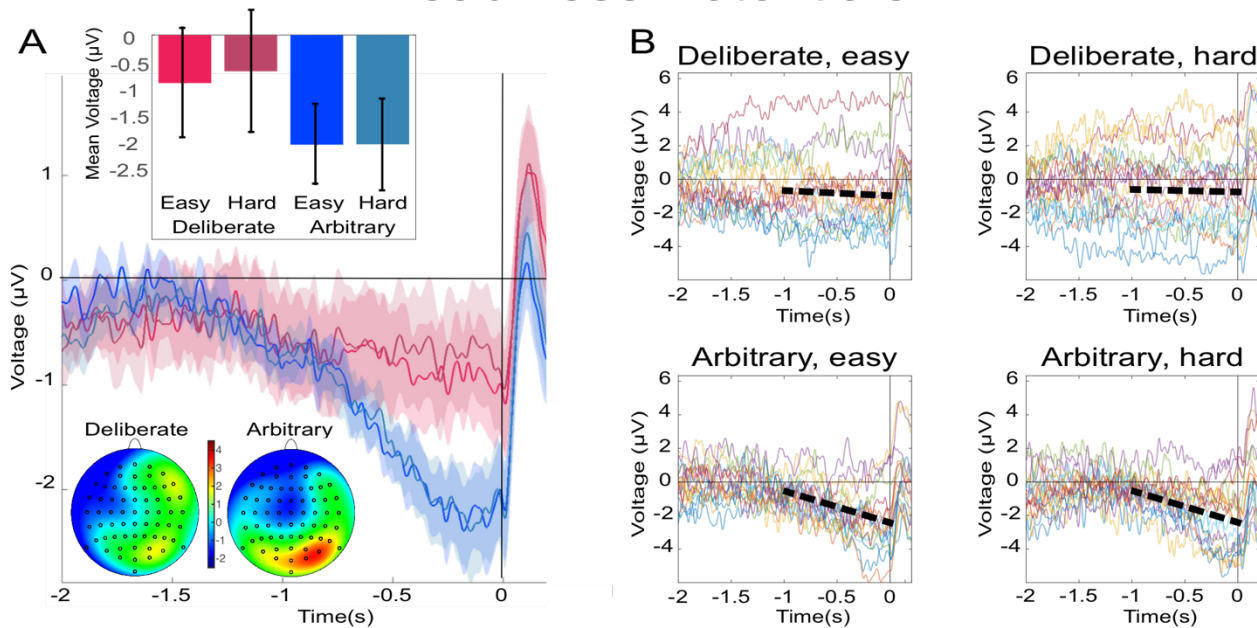


Figure 2: The readiness potentials for deliberate and arbitrary decisions. (A) Mean and SE of the Readiness Potential (RP) in deliberate (red shades) and arbitrary (blue shades) easy and hard decisions in electrode Cz, as well as scalp distributions. Zero refers to time of right/left movement, or response, made by the subject. Notably, the RP significantly differs from zero and displays a typical scalp distribution for arbitrary decisions only. The scalp distribution was calculated over the averaged activity during the last 500 ms before response, across subjects. The inset shows the mean amplitude of the RP, with 95% confidence intervals over the same time window. Response-locked potentials with an expanded timecourse, and stimulus-locked potentials are given in Fig. 6B and 6A, respectively. The same (response-locked) potentials as here, but with a *movement-locked baseline* of -1 to -0.5 s (same as in our Bayesian analysis below), are given in Fig. 6C. **(B)** Individual subjects' Cz activity in the four conditions (n=18). The linear-regression line for voltage against time over the last 1000 ms before response onset is designated by a dashed, black line. Note that the waveforms converge to an RP only in arbitrary decisions.

The second novel paradigm focused on personally meaningful choices. Subjects rated the favorability of various drinks in the first part of the experiment. Then they selected between pairs of those drinks in the second part of the experiment. In the deliberate-decision condition, subjects' selections determined the drinks they needed to drink at the end of each block and at the end of the experiment. For arbitrary decisions, subjects' choices had no influence over the drinks they had to drink. This was done by either letting subjects know that they would need to drink both drinks in arbitrary blocks, regardless of the button they pressed (arbitrary-different trials), or by presenting them with two identical drinks (arbitrary-same trials).

Like any new paradigm, we needed to rigorously test it behaviorally before we could begin EEG studies. We thus programmed this paradigm and ran it behaviorally with common store-bought drinks (in the United States) like apple juice, cranberry juice, ice tea, etc. (N=20). Subjects expressed strong preferences between the drinks in the first part of the experiment. However, an analysis of their reaction time (RT) and the consistency of their selections in the second part of the experiment with their selections in the first part of the experiment convinced us that the subjects did not in fact care that much which of the drinks they would need to drink. Post-experimental briefings provided further evidence for this.

Therefore, we swapped out many common drinks for ones that subjects would be likely to find aversive, like diluted soy sauce or diluted wasabi (reminiscent of horseradish). We also gave the subjects longer training blocks to be more sure that they completely understood the task before they began the study. This finally yielded the behavioral results we expected (N=10). Subject's reaction times were generally higher for deliberate than for arbitrary decisions and subjects' consistency was over 90%, on average, for deliberate trials and less than 65%, on average, for arbitrary trials.

We now had to validate this paradigm for a Libet-like measurement of decision-onset time to calibrate our results against those common in the literature. Given the strong criticism against the analogue Libet clock in the recent literature (Banks & Isham, 2011; Soon et al., 2008), we opted to use random letters that flip at constant intervals instead of an analogue clock. The letter stream would stop between 3 and 6 letters after the subjects pressed the left or right key. The subjects thus had to report what the letter was on the screen when they decided rather than where the spot was on the analogue clock (N=12). Unfortunately, we found that the letters had to flip every 500 ms or so for the subjects to carry out the experiment and time their decisions. For shorter durations, our subjects consistently reported not being able to tell apart the letters while concentrating on which drink they preferred.

So, we switched to consecutive letters that run in a loop (C, ..., Y, Z, A, B, C, ...). Only the first letter in the series was chosen randomly. This allowed us to bring the letter-flipping interval down to 200 ms, which we found more acceptable. Previous research suggests that this method is a viable one to measure decision onset (Banks & Isham, 2011) (N=12). One problem that remained was that subjects selected the last letter of the letter stream as the one that was on when they made

up their mind too often, even though it appeared up to a second after their movement onset. We therefore appended an # character at the end of the letter stream and informed subjects that they could not report the # as the symbol that was on when they decided. This finally led to clear and consistent results that were in line with the Libet literature (Banks & Isham, 2011; Haggard & Eimer, 1999; Libet, 1985; Libet et al., 1983).

Having run 30 subjects in this final version of the paradigm, we found that the choice consistency was high for deliberate decisions and much lower for arbitrary-different decisions, as expected (Figure 3). We found also found divergent W times between arbitrary and deliberate decisions (Figure 4), which cannot be simply due to reaction-time differences (Figure 3, middle), because the pattern in the W differences is other than that in the reaction-time difference. This surprising result casts further doubt on the decision-onset times reported in the Libet paradigm. So, we decided that it is worthwhile pursuing on its own right, while we continue to work on the paradigms proposed in the Bial proposal. We are therefore now running various followup control studies for the above.

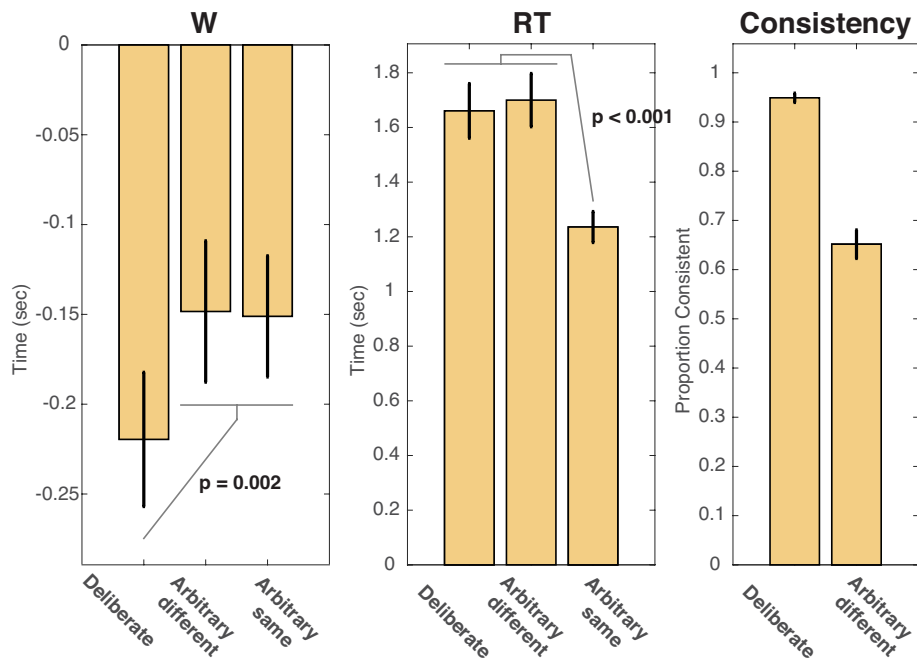


Figure 3: (Left) The bar graph depicts the length of time that passed between what subjects experienced as the onset of their decisions and the time of the button press in all 3 conditions. Differences between deliberate and the other conditions are significant, $p=0.002$ 1-way ANOVA. (Middle) Subject's reaction times in all 3 conditions. Differences between arbitrary-same and the other conditions are significant, $p=1 \cdot 10^{-10}$ 1-way ANOVA. (Right) Subjects' consistency with their initial ratings in the deliberate and arbitrary-different conditions. Consistency cannot be meaningfully defined or computed for the arbitrary-same condition. Panels A-C are all averages across all subjects (mean \pm s.e.m.).

It is known already from the original Libet experiments (Libet, 1985; Libet et al., 1983) that subjects perceive their movement onset (typically called M time) to roughly 85 ms before the actual movement onset. We wanted to know whether this error is modified with the type of decision that subjects make. We also wanted to check whether subjects' perception of stimulus onset (S time) was modified by the decision type. We therefore ran another paradigm very similar to the one above, yet where subjects had to either time the movement onset or the onset of the stimulus (N=30). For stimulus timing, we found that subjects made only small errors, S being roughly 15 ms after stimulus onset on average, with no significant differences between conditions ($p=0.3$). But for M time, subjects generally had a similar pattern to the one we observed for W times. The combination of the average W time and M time results can be found in Figure 5.

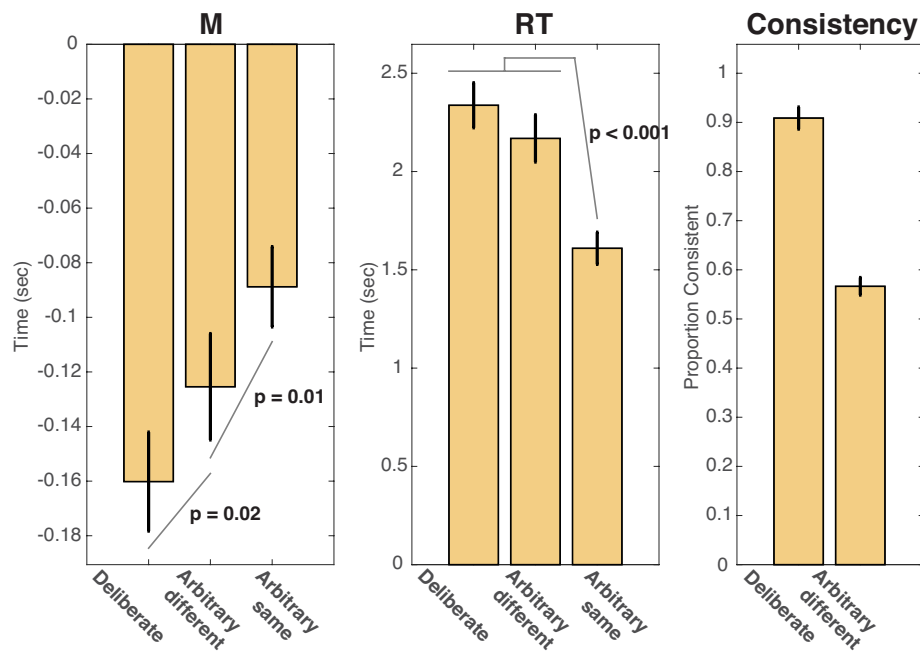


Figure 4: (Left) The bar graph depicts the length of time that passed between what subjects experienced as the onset of their button pressed and the time of the button press in all 3 conditions. Differences between deliberate and the arbitrary same as well as arbitrary same and arbitrary different are significant, $p=0.02$ and $p=0.01$, respectively, 1-way ANOVA. (Middle) Subject's reaction times in all 3 conditions. Differences between arbitrary-same and the other conditions are significant, $p=5 \cdot 10^{-8}$ 1-way ANOVA. (Right) Subjects' consistency with their initial ratings in the deliberate and arbitrary-different conditions. Consistency cannot be meaningfully defined or computed for the arbitrary-same condition. Panels A-C are all averages across all subjects (mean \pm s.e.m.).

We see that in all 3 decision conditions W time falls something like 40-70 ms behind M time. These results are in line with W time being backward inferred from M time rather than an independent mental event (Banks & Isham, 2009, 2011; Lau, Rogers, & Passingham, 2007).

For the above experiments, the W and M times were collected from different groups. So, within subject comparisons between W and M times cannot be made. We therefore now gathered M and W times within the same subject group. This will serve both as an additional replication of our results as well as let us directly compare these two timings within subjects. Furthermore, after calibrated and validating the drinks-selection paradigm against the Libet literature, we were able to gather EEG data from 24 subjects using that paradigm. And we are currently analyzing the data we gathered.

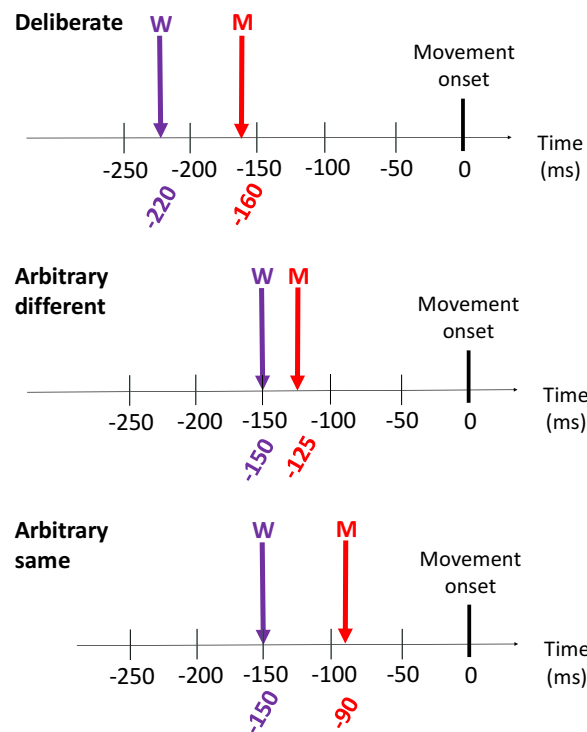


Figure 5: A visual depiction of the average W and M time for all 3 conditions,

In a fourth study, we first find the location and direction where a transcranial magnetic stimulation (TMS) pulse results in a twitch of a subject’s index finger. We then trigger that TMS pulse by electromyography (EMG) from the muscle that controls the twitching of the index finger. Finally, subjects are instructed to twitch that index finger at will within this closed-loop design of ours. Subjects typically report hearing the TMS click and their finger moving just as they were “about to decide to move”. We are currently collecting more data with this paradigm.

Last, we are making steady progress on further refining the online, real-time action-prediction system using EEG recordings—our fifth study. We anticipate being able to report statistics on collected data within a few months.

References

- Banks, W. P., & Isham, E. A. (2009). We infer rather than perceive the moment we decided to act. *Psychological Science*, *20*(1), 17.
- Banks, W. P., & Isham, E. A. (2011). Do We Really Know What We Are Doing? Implications of Reported Time of Decision for Theories of Volition. In W. Sinnott-Armstrong & L. Nadel (Eds.), *Conscious will and responsibility* (pp. 47-60): Oxford University Press.
- Fried, I., Mukamel, R., & Kreiman, G. (2011). Internally generated preactivation of single neurons in human medial frontal cortex predicts volition. *Neuron*, *69*, 548-562.
- Haggard, P., & Eimer, M. (1999). On the relation between brain potentials and the awareness of voluntary movements. *Experimental Brain Research*, *126*(1), 128-133.
- Harris, S. (2012). *Free will*. New York, NY: Simon & Schuster, Inc.
- Lau, H., Rogers, R., & Passingham, R. (2007). Manipulating the experienced onset of intention after action execution. *Journal of cognitive neuroscience*, *19*(1), 81-90.
- Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and brain sciences*, *8*, 529-539.
- Libet, B., Gleason, C., Wright, E., & Pearl, D. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential): The unconscious initiation of a freely voluntary act. *Brain*, *106*(3), 623.
- Maoz, U., Rutishauser, U., Kim, S., Cai, X., Lee, D., & Koch, C. (2013). Predeliberation activity in prefrontal cortex and striatum and the prediction of subsequent value judgment. *Frontiers in neuroscience*, *7*, 225.
- Maoz, U., & Yaffe, G. (2013). Cognitive Neuroscience and Criminal Responsibility. In M. Gazzaniga (Ed.), *Cognitive Neuroscience: The Biology of the Mind (Fifth Edition)*. Cambridge, MA: MIT Press.
- Mele, A. (2006). *Free will and luck*: Oxford University Press.
- Mele, A. (2009). *Effective intentions: the power of conscious will*: Oxford University Press, USA.
- Newell, B. R., & Shanks, D. R. (2014). Unconscious influences on decision making: A critical review. *Behavioral and brain sciences*, *37*(01), 1-19.
- Roskies, A. (2010). How Does Neuroscience Affect Our Conception of Volition? *Annual Review of Neuroscience*, *33*, 109-130.
- Sinnott-Armstrong, W., & Nadel, L. (2011). Introduction. In W. Sinnott-Armstrong & L. Nadel (Eds.), *Conscious will and responsibility: A tribute to Benjamin Libet*
Oxford Series in Neuroscience, Law and Responsibility: Oxford University Press.
- Soon, C., Brass, M., Heinze, H., & Haynes, J. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, *11*(5), 543-545.
- Soon, C., He, A., Bode, S., & Haynes, J. (2013). Predicting free choices for abstract intentions. *Proceedings of the National Academy of Sciences*, *110*(15), 6217-6222.
- Wegner, D. (2002). *The illusion of conscious will*: MIT Press.