Emulating Future Neurotechnology Using Magic

Jay A. Olson
Mariève Cyr
Despina Z. Artenie
Thomas Strandberg
Lars Hall

See next page for additional authors

Follow this and additional works at: https://digitalcommons.chapman.edu/psychology_articles

Part of the Artificial Intelligence and Robotics Commons, Bioelectrical and Neuroengineering Commons, Bioethics and Medical Ethics Commons, Medical Humanities Commons, and the Neurosciences Commons
Emulating Future Neurotechnology Using Magic

Comments
This article was originally published in Consciousness and Cognition, volume 107, in 2023. https://doi.org/10.1016/j.concog.2022.103450

Creative Commons License
This work is licensed under a Creative Commons Attribution 4.0 License.

Copyright
The authors

Authors
Jay A. Olson, Mariève Cyr, Despina Z. Artenie, Thomas Strandberg, Lars Hall, Matthew L. Tompkins, Amir Raz, and Petter Johansson
Emulating future neurotechnology using magic

Jay A. Olson a, *, Mariève Cyr b, Despina Z. Artenie c, Thomas Strandberg d, Lars Hall d, Matthew L. Tompkins d, Amir Raz e, 1, Petter Johansson d, *, 1

a Department of Psychology, McGill University, 2001 McGill College Ave., Montreal, QC H3A 1G1, Canada
b Faculty of Medicine and Health Sciences, McGill University, 3605 De la Montagne St., Montreal, QC H3G 2M1, Canada
c Department of Psychology, Université du Québec à Montréal, 100 Sherbrooke St. W., Montreal, QC H2X 3P2, Canada
d Lund University Cognitive Science, Lund University, Box 192, S-221 00, Lund, Sweden
e Institute for Interdisciplinary Behavioral and Brain Sciences, Chapman University, 9401 Jeronimo Road, Irvine, CA 92618, USA

ARTICLE INFO

Keywords:
Neurotechnology
Artificial intelligence
Neuroethics
Attitudes
Magic
Deception

ABSTRACT

Recent developments in neuroscience and artificial intelligence have allowed machines to decode mental processes with growing accuracy. Neuroethicists have speculated that perfecting these technologies may result in reactions ranging from an invasion of privacy to an increase in self-understanding. Yet, evaluating these predictions is difficult given that people are poor at forecasting their reactions. To address this, we developed a paradigm using elements of performance magic to emulate future neurotechnologies. We led 59 participants to believe that a (sham) neurotechnological machine could infer their preferences, detect their errors, and reveal their deep-seated attitudes. The machine gave participants randomly assigned positive or negative feedback about their brain’s supposed attitudes towards charity. Around 80% of participants in both groups provided rationalisations for this feedback, which shifted their attitudes in the manipulated direction but did not influence donation behaviour. Our paradigm reveals how people may respond to prospective neurotechnologies, which may inform neuroethical frameworks.

1. Introduction

1.1. Advances in neurotechnology

Novelist Arthur C. Clarke (2013) famously asserted that “any sufficiently advanced technology is indistinguishable from magic”. But the reverse can also be true: magic tricks can be made indistinguishable from advanced technology. When paired with real scientific equipment, magic techniques can create compelling illusions that allow people to experience prospective technologies first-hand. Here, we demonstrate that a magic-based paradigm may be particularly useful to emulate neurotechnologies and potentially inform neuroethical frameworks.

Broadly defined, neurotechnology involves invasive or non-invasive methods to monitor or modulate brain activity (Goering et al., 2021). Recent developments in neural decoding and artificial intelligence have made it possible, in a limited fashion, to infer various aspects of human thought (Ritchie et al., 2019). The pairing of neural imaging with machine learning has allowed researchers to...
decode participants’ brain activity in order to infer what they are seeing, imagining, or even dreaming (Horikawa et al., 2013; Horikawa and Kamitani, 2017). For example, one study identified the neural correlates of viewing various face stimuli; EEG data from a single participant could be used to determine which of over one hundred faces was being presented (Nemrodov et al., 2018). Other studies have used fMRI brain activity patterns to infer basic personality traits after exposing people to threatening stimuli (Fernandes et al., 2017). Similar decoding methods have also been used to determine what verbal utterances participants were thinking about in real time (Moses et al., 2019).

Other recent developments have enabled researchers to decode information that participants are not even aware of themselves. One fMRI study decoded the semantic category of words (e.g., animal or non-animal) presented below the level of awareness (Sheikh et al., 2019). Researchers have used the same method to infer which of two images participants would choose several seconds before the participants themselves were aware of making this decision (Koenig-Robert and Pearson, 2019).

Although these findings are impressive, brain reading remains in its infancy. The information decoded from brain activity is often relatively rudimentary and requires cooperation from participants. Brain reading is further limited by the cost and technical expertise required to design and operate the imaging machines. Nevertheless, given that brain reading has the potential to become a powerful and commonplace technology in the future (Yuste et al., 2017), it is important to avoid the delay fallacy wherein discussions of the implications of emerging technologies lag behind the technological frontier (Mecacci and Haselager, 2017; van de Poel and Royakkers, 2011).

1.2. Ethical implications

Ethicists have accordingly started to speculate about the potential ramifications of various neurotechnologies. Future developments in neural decoding may carry implications across several domains including personal responsibility, autonomy, and identity (Goering et al., 2021; Ryberg, 2017). For example, brain reading could be used to predict the risk of recidivism (Ienca and Andorno, 2017) or to influence attributions of criminal responsibility by inferring one’s mental state at the time of the crime (Meynen, 2020). Regarding autonomy, employers could use future brain reading to screen out undesirable characteristics in their employees. Brain reading also has the potential to undermine personal identity by changing how we think about ourselves. Some people may see feedback from neurotechnology as a more objective and accurate representation of personality traits, biases, or beliefs than those accessible through introspection (cf. Berent and Platt, 2021). In this way, technology may trump our subjective experiences in the understanding of who we are.

Although neurotechnology could potentially boost self-understanding, many people find the prospect of brain reading intrusive (Richmond, 2012); it violates the long-standing expectation that one’s thoughts are private (Moore, 2016). The implications of this potential loss of privacy, however, remain unclear. Thomas Nagel (1998, p. 4) argues that such privacy is fundamental to a properly functioning society: “the boundary between what we reveal and what we do not, and some control over that boundary, are among the most important attributes of our humanity.” Conversely, aside from nefarious uses such as government control, Lippert-Rasmussen (2016) argues that access to others’ thoughts could offer an additional source of information to foster intimacy and authenticity. In his view, “the gaze of others would become much less oppressive if everyone’s inner lives were transparent to everyone else” (p. 230). The speculated consequences of future neurotechnology thus show considerable range.

Importantly, these consequences may not remain merely speculative. Given the widespread and complex implications of future brain reading technologies, ethicists have proposed forward-thinking policies such as the adoption of “neurorights” to protect citizens (Baselga-Garriga et al., 2022; Yuste et al., 2017). These efforts to safeguard people from the uses and misuses of brain reading depend, in part, on our ability to anticipate people’s future reactions. More caution is needed, for example, if people see brain reading as an invasion of privacy versus a novel way to promote authenticity.

However, simply asking people about how they would react to future neurotechnologies may be insufficient. People often overestimate their responses to future events (Dillard et al., 2020; Gilbert et al., 1998) and have difficulty explaining their attitudes reliably (Hall et al., 2012). One study found that when people read vignettes of neurotechnology predicting and influencing behaviour, they interpret the situations based on their current metaphysical assumptions, even if these assumptions would be contradicted by the information in the vignettes (Rose et al., 2015). Reasoning hypothetically about a future machine may have limited validity compared to the concrete experience of having a machine control one’s mind. Instead, “Wizard of Oz” prototyping could offer a potential solution (Kelley, 1984). Here, a simulation is created of a future product by fabricating an apparently working prototype, which is then tested in real-world scenarios to generate more accurate responses from users.

1.3. Magical neurotechnology

We developed a Wizard of Oz-style paradigm to emulate prospective neurotechnologies based on elements of performance magic. Indeed, many of the abilities enabled by future neurotechnologies can be mimicked using magic tricks. Most relevant is the branch of performance magic known as mentalism, which involves mimicking abilities such as mind reading, thought insertion, and prediction. A brain scanner decoding a participant’s thoughts resembles a magician reading the mind of a spectator, and a device that inserts thoughts to affect behaviour resembles magicians influencing the audience’s decisions without their awareness (Olson et al., 2015). In this way, magic could create the compelling illusion of future neurotechnological developments before they are available.

We have previously demonstrated the believability of combining magic with neurotechnology by convincing university students that a brain scanner could both read and influence their thoughts (Olson et al., 2016). In a condition designed to simulate mind reading, participants chose an arbitrary two-digit number while inside a sham MRI scanner. The machine ostensibly decoded their
brain activity while they focused on the number. A simple magic trick allowed the experimenter to demonstrate that the machine’s decoded number matched the one that the participant had previously chosen. The same magic trick was then used to simulate thought insertion. In this mind-influencing condition, participants were again instructed to think of a number. Instead of being told that the machine would decode their brain activity, they were told that the machine would manipulate their brain through “electromagnetic fluctuations”. The magic trick made it appear as if the machine had randomly chosen a number and then influenced participants to choose it. In this condition, participants felt less control over their decisions and reported a range of experiences, including hearing an ominous voice controlling their choices.

By combining neuroscientific-looking props with magic, we were thus able to convince educated participants to both believe in and directly experience a “future” machine that could accurately read and influence their decisions. However, given the relatively inconsequential target of the brain reading — arbitrary number choices — it is difficult to assess how participants would react to having the machine decode thoughts that are more meaningful or private, including those relevant to neuroethics.

1.4. Present study

Here, we extend our method to create a future context in which brain reading is powerful enough to decode information central to the self, such as political attitudes. We focused on attitudes towards charity because people often believe that such moral values characterise one’s “true self” (Strohminger et al., 2017). According to the lay understanding, this true self is a more private and accurate version of the self that is indicative of one’s core identity (Schlegel et al., 2011). We aimed to manipulate this core aspect of the

![Fig. 1. Overview of the experimental procedure. Leveraging three magic tricks, the machine appeared to decode participants’ preferences, detect their written errors, and reveal their brain’s deep-seated helping attitudes.](image-url)
self in order to assess reactions to more personal and ethically relevant domains. To do so, we emulated a neurotechnological machine that could identify people’s attitudes towards charity better than their own introspection. First, we aimed to explore how participants would react to a potential invasion of mental privacy by having a machine seemingly infer their consumer preferences and political attitudes. Second, we explored the crucial issue of people’s trust in neurotechnology by simulating a scenario in which the machine could give personal feedback that is inconsistent with what participants report. Finally, we investigated how people might adapt their own beliefs based on this discrepant feedback. How might people react to this dissonance between their own subjective feelings and the machine’s seemingly objective assessment? Could such brain reading supersede one’s own judgement? We present a novel method to begin answering these questions.

2. Methods

2.1. Participants

We recruited 62 participants from social media postings to complete a study ostensibly testing a new type of neurotechnology. Three participants expressed discomfort with entering the sham MRI as part of the experimental procedure and did not begin the study; the remaining 59 participants were on average 21.7 years old (SD = 4.0) and 71% were female. Most were undergraduate students (80%), commonly in the third year of their studies. Others were graduate students (15%) and one was a post-doctoral fellow. We pre-screened participants to exclude those in psychology who may have heard about our previous magic-based studies.

We chose a sample size of 60 participants in advance based on feasibility. In line with our previous study showing how a similar sham machine can manipulate feelings of control over thoughts (Olson et al., 2016), we expected large effects (d ≈ 1). Using a non-directional independent-samples t test as an estimate, this sample size would give us 97% power to detect differences of one standard deviation between the two groups in our design.

2.2. Procedure

We aimed to investigate whether participants would believe that a (sham) neurotechnological system could accurately read their thoughts and attitudes. The system included a sham MRI scanner and an EEG system that supposedly used neural decoding driven by artificial intelligence (AI). The procedure had three main phases: basic brain reading, human error detection, and attitude feedback (Fig. 1). In each phase, participants were asked to think about their agreement with various preferences or attitudes while the sham EEG system appeared to accurately decode their answers. We combined several layers of subtle yet elaborate deception to make the procedure more believable (cf. Olson and Raz, 2021). The study was approved by the McGill University Research Ethics Board II (#436–0416).

2.2.1. Briefing

Participants were told that the goal of the study was to build a database of neural activation patterns associated with various political attitudes. At the start of the procedure, the experimenter elaborated on the fictitious premises of the study, explaining that people have two kinds of attitudes: stated attitudes, which people believe they hold, and true attitudes, which people hold unconsciously.

---

Fig. 2. After briefly entering the MRI scanner, participants put on an EEG cap and thought about their agreement with various attitudes while the machine appeared to infer and print out their responses.
in their brain and can be decoded from neural activity. The participants were told that for simple topics, such as basic consumer preferences, stated and true attitudes tend to align; but for more complex topics, such as political views, these types of attitudes may diverge. The experimenter explained that when they diverge, the true attitudes (as measured by the machine) tend to better predict behaviour than the stated ones.

After giving informed consent, participants completed demographic and mood questionnaires (see Measures). They then completed an MRI safety screening and entered a plausible-looking though sham scanner (Psychology Software Tools, Inc., Sharpsburg, PA) which ostensibly recorded their brain structure to assist with neural decoding. The room also contained an EEG system and various scientific-looking props (Fig. 2) (cf. Bernstein et al., 2020).

2.2.2. Phase 1: Simulating brain reading of preferences and attitudes

To begin our hoax scenario, we intended to build participants’ trust in the machine by pretending that it could decode their preferences and attitudes. We chose consumer preferences as a suitable initial domain because participants already have some familiarity with AI-driven recommendation systems for various products.

The experimenter fitted participants with an EEG cap that would supposedly read their preferences and attitudes. Each trial, the participants would see two stimuli (e.g., vanilla versus chocolate ice cream) and would think of a number from 0 to 100 corresponding to their preference on a visual analogue scale. Participants would silently concentrate on the number for five seconds until a printer would print a sheet of technical-looking (but bogus) output. The experimenter would look at the computer screen, scribble down the number that the machine ostensibly inferred from the participant’s neural activity, and then ask the participant to state the chosen number. Using a magic trick (see Supplementary methods), the machine’s inference matched the participant’s number. For subtlety, the experimenter did not mention that the numbers matched; he merely asked the participants to write down their chosen number beside the machine’s inference so that the participants would notice the match on their own (cf. Olson and Raz, 2021). To make the machine decoding more believable, only on the first trial, the output number was incorrect (off by 10); such fake mistakes can make mentalism performances appear more realistic (Olson and Raz, 2021; Pailhès et al., 2022). The subsequent trials compared apples versus oranges and soft drinks versus smoothies. After the three trials, the experimenter stated that the machine was calibrated and they could continue with more abstract attitudes.

We chose political attitudes as the next domain because they are more personal than consumer preferences yet people are still used to having them assessed (e.g., in opinion polls). Our survey included five arbitrarily chosen political questions on issues such as environmentalism and government affairs (e.g., “The government should impose stricter gun control”). As in the previous phase, participants would read each statement and then think of their numerical agreement before writing it down on the survey. This time, now that the machine was apparently calibrated, we did not show participants the immediate feedback.

We used a different method to have the machine infer the participants’ responses; mixing methods can obscure the secret behind magic tricks (Masters, Bagienski, Kuhn, & Smith, in progress). This time, the secret involved a research assistant who discreetly observed the participants’ written answers through a one-way mirror. We created the illusion that the machine was printing out the chosen number before the participants wrote it down (see Supplementary methods). When the experimenter showed the participants the printed machine feedback at the end, it appeared as if the machine correctly inferred their responses.

2.2.3. Phase 2: Simulating human error detection

Thus far, participants experienced brain reading that was congruent with their expectations: the machines’ output matched their reported attitudes. For the next phase, we were interested in investigating how participants might react when the machine’s output is incongruent. This discrepancy could be due to insightful AI, which is able to decode information that people are unaware of or that transcends the biases and limitations of human judgement. Or, the AI may be erroneous due to malfunction, bias, or even malevolence on the part of the designer or an adversary. Here, we aimed to further increase trust in the machine by making it appear as if it knew the participants’ attitudes better than they did themselves.

To do so, we orchestrated a situation in which participants appeared to make an error in their written report, which the machine would then detect and correct. This ruse was accomplished by leveraging change blindness, the phenomenon wherein people fail to notice changes in their visual field (e.g., Simons and Levin, 1997). The assistant behind the one-way mirror filled out a duplicate version of the participant’s political attitude questionnaire, with all of the same responses except for one intentional error. Namely, on a question about whether the government should monitor phone calls to prevent terrorism, the assistant marked down the reverse of what the participant wrote. For example, if the participant marked 25% agreement, the assistant would mark 75% agreement instead.

The experimenter switched the participant’s questionnaire with this erroneous duplicate. As in other studies (Hall et al., 2013; Strandberg et al., 2020), none of the participants spontaneously mentioned noticing this discrepancy. The participants thus had a questionnaire that exactly matched their original answers, except for this one mismatch. Because participants failed to detect the switch, we were able to get them to attribute the discrepancy between their intended answer and their written answer as their own error. (See Supplementary methods for an explanation of the questionnaire switch.)

The experimenter checked the printed output described in the previous phase. He flipped through each print-out, and when he reached the manipulated answer regarding phone monitoring, he commented, “It looks like there’s a discrepancy here” (between the written answer and the machine output). In response, 93% of the participants stated that they had made an error and corrected the answer on the questionnaire. When the experimenter asked this group to confirm which of the two ratings reflected their true attitude, all indicated that the machine’s output was accurate. As one of them stated, “The machine was right — I was wrong”. The 7% of participants who did not mention the error generally acted shy and confused.
2.2.4. Phase 3: Simulating insight into attitudes

Having established more trust in the machine, we could approach the question of what would happen when the AI provides insight beyond simple error detection. We centred the investigation on moral values such as charity, which people often characterise as part of their “true self” or “core identity” (Schlegel and Hicks, 2011; Strohminger et al., 2017). Using a similar setup as in the previous phase, participants completed the Attitudes Towards Helping Others scale (see Measures; Webb et al., 2000). To help conceal our main measure and reduce demand characteristics, the helping items were intermixed with arbitrarily selected filler items: four relating to the environment (e.g., “The state of the natural environment has a large effect on my well-being”) and four on nationalism (“Countries should invest more money in military defence”).

After completing the questionnaires, the participants were randomly assigned to receive either positive or negative false feedback from the machine. The randomisation had the provisions that gender should balance between the conditions and that those with an average over 90% on the helping items would be assigned to the negative group, in order to perform the upcoming manipulation. This latter provision meant that the negative feedback group started out with slightly higher helping attitudes, by 7 points out of 100 ($t(57) = 2.15, p = .036$). This difference did not change any outcomes of our hypothesis tests.

As in the previous phase, the assistant created an output that was apparently generated by the machine. This output matched the answers provided by the participant, with the exception that three of the four helping items were shifted in the direction of the false feedback. The positive group had these three items increased to an average of 82% agreement; the negative group had the same items reduced to 32% agreement. The experimenter went through each manipulated item and asked the participant, “Do you know why there would be a discrepancy here? Can you think of times in your life in which you behaved more consistently with your true attitude [as the machine guessed] than your stated one [as you wrote]?” We tape recorded and transcribed their verbal responses, which two independent judges later coded based on whether the participants (1) agreed with and justified the machine feedback, or whether they (2) disagreed with the results or could not explain them. This dichotomous distinction was clear: the judges had 100% reliability in their coding. Another two judges later performed an exploratory thematic analysis to check for patterns in the content of the responses; one judge generated themes before both categorised each response.

After leaving the experiment room, the experimenter gave participants a final questionnaire and (falsely) stated that the study had completed. The questionnaire contained the Attitudes Towards Helping Others scale intermixed among the previous filler items as well as an additional two (e.g., “Governments, rather than students, should pay tuition fees”). They also completed the same mood questionnaire as before (see Measures). These questions allowed us to assess whether the feedback influenced participants’ attitudes or mood.

2.2.5. Donation behaviour

Finally, we measured the possible effect on helping behaviour. The experimenter gave a partial debriefing (that did not reveal the deception) and then compensated the participant with $10 in coins. The participant left the lab and returned to the building’s lobby, which had a donation table set up by the door. An assistant at the table asked whether the participant would like to make a donation for clinical research on mental health. We then contacted participants by email with a full debriefing of the study and offered to return their donated money; none requested it back so we donated it to a relevant charity.2

2.3. Measures

2.3.1. Helping attitudes

Our main dependent variable was the Attitudes Towards Helping Others scale (Webb et al., 2000). The scale contains four items, such as “People should be willing to help others who are less fortunate” and “Helping troubled people with their problems is very important to me”. We adapted the response format to a visual analogue scale to match the other questions; participants rated their agreement from 0 to 100. Reliability was acceptable to high before (Cronbach’s $\alpha = .75$) and after ($\alpha = .92$) receiving the machine feedback.

2.3.2. Mood

We used the Positive and Negative Affect Schedule (I-PANAS-SF) to measure mood (Thompson, 2007). Participants rated the extent to which they were feeling 10 positive (e.g., Excited) and 10 negative emotions (e.g., Afraid) on a scale ranging from 1 to 5. Reliability was good at both pre- and post-test for positive ($\alpha = .83 \text{ .85}$) and negative affect ($\alpha = .86, .80$).

2.4. Analysis

We used mixed-effect modelling to predict the average helping attitude (i.e., the Attitudes Towards Helping Others score) given the condition (positive or negative feedback), time (pre- or post-feedback), and the interaction, with a random intercept for each participant. We used the same method to check for changes in both positive and negative affect. To assess differences in donation amounts between the conditions, we used a Wilcoxon–Mann–Whitney test given that normality assumptions were not met. All tests were non-directional, used a Type I error rate of .05, and had no family-wise error correction. Square brackets throughout denote 95%

---

2 We initially had one additional dependent variable to assess helping behaviour: whether participants picked up a lost key by the building door. This measure proved unreliable; several participants did not notice the key and strangers would often pick it up first. We thus excluded this measure.
confidence intervals.

3. Results

3.1. Participants rationalised the machine feedback

As the machine seemingly inferred participants’ preferences and attitudes, many expressed amazement by laughing or calling it “cool” or “interesting”. For example, one asked, “Whoa, the machine inferred this? … Oh my god how do you do that? Can we do more of these?” None of the participants voiced any suspicion about the mind-reading abilities of the machine throughout the study.

After receiving the randomly assigned feedback on charity attitudes, the large majority of participants agreed with it and justified its accuracy (positive: 79% [66%, 93%]; negative: 80% [63%, 93%]). For example, some participants stated that the machine’s feedback showed what they felt “in [their] heart” or that the machine’s assessments “[were] all true”. Some found the process of receiving the feedback “stunning”, “demoralising”, or “enlightening”; at the extreme, one participant stated that he was grateful that the machine feedback helped him realise how much his life had diverged from his ideals. One participant who received negative feedback explained:

I feel that recently there’s been a shift in me that I feel that people should help themselves, but that sort of goes against how I’ve been raised and all my friends around me. … I really felt this change within me from [university]. When I started [university] I was part of [a socialist group] and I was really into student politics. But now … I feel like I’ve completely changed.

In contrast, one who received positive feedback stated:

I think I would imagine that, [the attitude feedback] would maybe be my ideal. But then when I think of the [response], I feel like maybe I calibrated it as, “Okay, in a perfect world, that’s what I would consider.” But then taking into account the scarcity of time and energy, I would expect less of people.

Between the experimental conditions, people generally discussed similar topics (cf. Johansson et al., 2006), such as homelessness and personal responsibility, but they differed in their explanations based on the feedback (see Table 1). Participants who did not provide a justification (20%) disagreed with or could not explain the discrepancy in the feedback.

3.2. The machine feedback influenced attitudes

We predicted that the machine feedback would shift participants’ subsequent reports of their charity attitudes. Consistent with this prediction, the positive group increased from 66.79 [62.77, 70.96] to 73.28 [68.12, 78.62], while the negative group decreased from 74 [69.37, 78.54]) to 61.58 [56.07, 67.69] (Fig. 3). Almost all participants (88%) changed their responses in the direction of the prediction, the positive group increased from 66.79 [62.77, 70.96] to 73.28 [68.12, 78.62], while the negative group decreased from 74 [69.37, 78.54]) to 61.58 [56.07, 67.69] (Fig. 3). Almost all participants (88%) changed their responses in the direction of the prediction. This resulted in a large standardised difference (Cohen’s $d$ = 3.34 [1.69, 3.05]) yet a more modest effect on raw difference scores in helping attitudes ($M_0 = 18.90 [14.76, 23.14]$ on a 100-point scale). See Table 2 for regression results.

3.3. Donation behaviour and mood did not change

Both groups donated on average around $1 (positive: $0.90 [0.57, 1.28]; negative: $1.06 [0.75, 1.39], Wilcoxon–Mann–Whitney $z = 0.94, p = .35$). There were no correlations between helping attitudes and donation amounts ($r < .25$). There was little change in

<table>
<thead>
<tr>
<th>Theme</th>
<th>Feedback</th>
<th>%</th>
<th>Example quote</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homelessness</td>
<td>Positive</td>
<td>8</td>
<td>I buy homeless people coffee a lot … if I have extra food I’ll give it to someone who needs it.</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>25</td>
<td>[I walk] past homeless people in the street every day and [I don’t feel] the need to help them out.</td>
</tr>
<tr>
<td>Donating</td>
<td>Positive</td>
<td>19</td>
<td>I donate whenever I can … Everything about this [feedback] is very true.</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>12</td>
<td>Right now [I’m] not able to donate money or help people economically.</td>
</tr>
<tr>
<td>Responsibility</td>
<td>Positive</td>
<td>10</td>
<td>I used to be more like “welfare users are taking advantage” but I’m realising this is most often not the case. So it is something I have not yet completely internalised but I am most definitely working on it.</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>20</td>
<td>I feel that recently there’s been a shift in me that I feel that people should help themselves, but that goes against how I’ve been raised and my friends around me.</td>
</tr>
<tr>
<td>Friends/family</td>
<td>Positive</td>
<td>10</td>
<td>Probably the way I was raised [explains the discrepancy in feedback].</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>10</td>
<td>[Sometimes] my friends [come] to me with their problems and [I don’t feel] any need to help them or [feel] annoyed … when I should be feeling supportive.</td>
</tr>
<tr>
<td>Practicality</td>
<td>Positive</td>
<td>15</td>
<td>I’m a pretty empathetic person, but … the practicality of … being really empathetic to other people is kind of more difficult in real life [which is why I initially rated my attitude lower].</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>3</td>
<td>I answered this because in principle it would [help], but practically I doubt it.</td>
</tr>
<tr>
<td>Priorities</td>
<td>Positive</td>
<td>5</td>
<td>Logic tells me that you shouldn’t be [as] helpful [as I am] but I have a good heart!</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>10</td>
<td>I was doing a quick conflict in my head between that classic awful conservative attitude of “charity for people who are wasting it and not working hard enough” versus “charity for people who [deserve it].”</td>
</tr>
</tbody>
</table>
Consciousness and Cognition 107 (2023) 103450

mood for both positive \( \beta = -0.34, t(55) = -1.65, p = .105 \) and negative affect \( \beta = 0.46, t(57) = 1.77, p = .082 \). The observed changes in attitude were thus unlikely due to differences in mood alone; mood changes did not correlate with attitude changes \( |r| < .14 \).

### 4. Discussion

We used magic and deception to emulate future neurotechnology that could decode mental content, detect errors, and provide insight into deep-seated attitudes. After the machine gave positive or negative feedback regarding their brain’s supposed attitudes towards charity, participants changed their attitudes in the congruent direction. The large majority of the participants (80%) agreed with this randomly assigned machine feedback and provided rationalisations to support it. However, we did not see changes in their donation behaviour measured immediately after the study.

The amount of agreement and the elaborateness of the rationalisations demonstrate the participants’ trust in brain-related feedback. Adding to the persuasiveness of our procedure, offering feedback about the brain leverages the popular understanding of the relationship between the mind and brain known as neuroessentialism. This viewpoint assumes that all experiences can be reduced to brain activity: the self is caused by the brain (Schultz, 2015), “a secular equivalent to the soul” (Racine et al., 2010, p. 728). When asked to rate the essence or core personality of an individual, for example, people put more weight on the results of brain tests rather than behavioural measures (Berent and Platt, 2021). Analogously, in our study, participants may have given more weight to the brain feedback than to their own introspection. The act of verbally justifying the discrepancies may have then helped to shift their attitudes (cf. Strandberg et al., 2018).

Participants’ justifications may have also helped the attitude change generalise from the three manipulated items on the helping questionnaire to the non-manipulated one (Fig. 3). Similar studies have shown related generalisation, such as false feedback about political attitudes influencing voting intentions (Hall et al., 2013). Our observed change in attitudes, however, did not extend to the donation measurement. One possible explanation for this null result was that the link between the attitude questionnaire and the target behaviour was somewhat weak. Rather than using the Attitudes Towards Helping Others Scale, we could have used the Attitudes Toward Charitable Organisations Scale (Webb et al., 2000). This scale may have been more relevant at the expense of potentially

---

**Table 2**

Standardised regression results. The feedback caused an attitude change in the congruent direction.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>( \beta )</th>
<th>CI</th>
<th>SE</th>
<th>( t )</th>
<th>( df )</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-0.14</td>
<td>-0.49 -0.21</td>
<td>0.18</td>
<td>4.32</td>
<td>57</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Time (post-feedback)</td>
<td>0.43</td>
<td>0.23 -0.63</td>
<td>0.10</td>
<td>4.32</td>
<td>67</td>
<td>0.058</td>
</tr>
<tr>
<td>Feedback (negative)</td>
<td>0.48</td>
<td>-0.02 -0.97</td>
<td>0.25</td>
<td>1.93</td>
<td>67</td>
<td>0.058</td>
</tr>
<tr>
<td>Interaction</td>
<td>-1.25</td>
<td>-1.53 -0.97</td>
<td>0.14</td>
<td>-8.98</td>
<td>57</td>
<td>&lt; .001</td>
</tr>
</tbody>
</table>

---

**Fig. 3.** Attitudes Towards Helping Others (visual analogue scale) agreement by condition and time. Participants adapted their attitudes to the false feedback they received. The first three listed items received manipulated feedback; the fourth did not. Dots show means, lines show changes over time, and error bars show 95% confidence intervals.
increasing demand characteristics; participants may have been more likely to suspect that the donation measurement was part of the study. Other studies have also found limited generalisation following false feedback (e.g., Merkelbach et al., 2011). The extent of generalisation in similar paradigms remains an open question (Artenie et al., under review).

4.1. Future studies and implications

Our methods could be used to evaluate how people will respond to future neurotechnologies. Given rapid advances in neuroscience and AI, researchers are speculating about the implications of future neurotechnologies and have proposed various competing hypotheses. Our paradigm allows for early empirical testing of these hypotheses which could help neuroethicists understand and predict their implications. As a proof of concept, in our study, we focused on questions relating to the acceptability of neurotechnology to decode private attitudes and about the malleability of the “true self” after receiving discrepant feedback from a neuroimaging machine. Some researchers have speculated that concepts such as the sense of self or free will are too complex and resilient to be heavily disrupted by neuroscientific information (de Cunha and Relvas, 2017; O’Connor and Joffe, 2013). However, there may be a difference between neuroscientific information and experience; the latter may have more potent effects. When faced with a scenario indistinguishable from advanced neurotechnology revealing insight about one’s brain, our findings suggest that people may adapt to feedback from a machine even though it conflicts with their own introspection.

Similar experimental studies using magic to emulate future neurotechnology could test other questions related to neuroethics. For example, where are the limits of privacy and comfort when a machine is reading one’s mind? Our participants expressed no issues with having a machine read their consumer preferences, political beliefs, and moral attitudes. Would they also be comfortable with a machine assessing their hidden biases, untold secrets, or embarrassing information? Could this method generate more accurate private information than typical self-report? Or, how would people react to a machine that, after a brief brain scan, was able to predict their future decisions? We suspect that people’s responses would differ when responding to hypothetical neuroscientific vignettes versus having compelling brain-reading experiences with seemingly genuine neurotechnology. Assessing such responses could help form a more empirical neuroethical framework by sketching out the boundaries of what people find acceptable when they actually experience it.

4.2. Limitations

Perhaps the largest limitation of our study was that we could not rule out demand characteristics, which could have influenced participants’ attitude reports after the manipulation. This is especially relevant given that we did not see a change in donation behaviour, which may have been less likely to be affected by demand characteristics because it carried a monetary cost and ostensibly occurred after the study. An indirect way to assess demand characteristics could have been to measure suspicion during the partial debriefing, but this may have roused suspicion (Olson and Raz, 2021) and invalidated our final donation measure. In any case, we at least saw no overt demonstrations of suspicion and found low rates in similar studies using sham machinery (Ali et al., 2014; Olson et al., 2016) and complex layers of deception (Olson and Raz, 2021). Further, the same-day debriefing, which was intended to minimise any effects on behaviour, prevented us from taking any follow-up measures to assess how long the attitude changes may have lasted; other studies have found that confabulating about false feedback can change attitudes for at least one week (Artenie et al., under review; Strandberg et al., 2018).

Another limitation has to do with the lack of a separate control group. Future studies could test our manipulation against a potentially weaker one, such as reading a vignette about receiving neurotechnology-based feedback. Or, the neurotechnology frame could be compared against receiving similar feedback from a sham personality test or a different class of technology. These control groups would help clarify the potency of receiving feedback from an emulated machine as well as the potential role of demand characteristics.

4.3. Conclusion

In sum, we believe that our admittedly uncommon and elaborate paradigm may help produce realistic reactions to future neurotechnologies. This paradigm offers promise in emulating these neurotechnologies to better understand and prepare for their eventual consequences.

5. Data statement

The dataset is available online (https://osf.io/5zgkd/).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
Acknowledgements

This study was funded by the BIAL Foundation (#102/2016). JO acknowledges the Joseph-Armand Bombardier Scholarship from the Social Sciences and Humanities Research Council of Canada, and the Fonds de recherche du Québec — Santé (FRQs). TS acknowledges the Uno Otterstedt Foundation (RFv 2016–0598) and Knut och Alice Wallenberg’s Foundation (RFh 2015–0406). MT was funded by the Crawford Foundation. AR acknowledges the Canada Research Chair program, Discovery and Discovery Acceleration Supplement grants from NSERC, and Canadian Institutes of Health Research. TS, LH, and PJ were funded by a Swedish Research Council grant (2014–137).

We would like to thank Kylar D’Aigle, Jason Da Silva Castanheira, and Ewalina Jeyanesan for transcribing data, as well as Denis Chmoulevitch, Johnny Nahas, Madalina Prostean, and Leah Suisa-Rocheleau for analysing the justifications.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.concog.2022.103450.

References


