Running Head: MORAL JUDGMENTS AND FREEDOM

Forthcoming in *Psychological Inquiry*

Moral Judgments and Intuitions about Freedom

Jonathan Phillips[1] and Joshua Knobe[1]

[1]*Department of Philosophy, University of North Carolina-Chapel Hill, Chapel Hill, North Carolina*

For correspondence:

Jonathan Phillips
Department of Philosophy
Caldwell Hall
UNC Chapel Hill
Chapel Hill, NC 27599

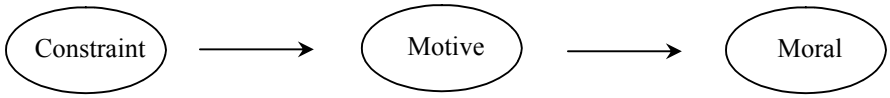
Email: philli@email.unc.edu

Reeder's article offers a new and intriguing approach to the study of people's ordinary understanding of freedom and constraint. On this approach, people use information about freedom and constraint as part of a quasi-scientific effort to make accurate inferences about an agent's motives. Their beliefs about the agent's motives then affect a wide variety of further psychological processes, including the process whereby they arrive at moral judgments.

In illustrating this new approach, Reeder cites an elegant study he conducted a number of years ago (Reeder & Spores, 1983). All subjects were given a vignette about a man who goes with his date to a pizza parlor and happens to come across a box that has been designated for charitable donations. In one condition, the man's date then requests that he make a donation; in the other, she requests that he steal the money that is already in the box. In both conditions, the man chooses to comply with this request. The key question is how subjects will use his behavior to make inferences about whether he is a morally good or morally bad person.

The results revealed a marked difference between conditions. When the man donated to charity, subjects were generally disinclined to conclude that he must have been a morally good person. It is as though they were thinking: 'He didn't just do this out of the goodness of his heart; he only did it because his date wanted him to.' By contrast, when the man stole the money, subjects tended not to discount on the basis of situational constraint. They had no problem concluding that he truly was an immoral person. As Reeder notes, a number of other studies have shown similar effects (McGraw, 1985, 1987; Trafimow & Trafimow, 1999; Vonk & van Knippenberg, 1994).

How are we to account for this phenomenon? Reeder proposes a model that looks more or less like this:



Here, people use facts about situational constraint to make inferences about the agent's motives, which in turn serve as input to the process through which they make moral judgments about the agent's character. Reeder's suggestion then is that this model can explain the effects observed in the experiment. Subjects would go through three stages:

(1) In both conditions, they use information from the story to infer that the agent is under situational constraint.

(2) They then use this information about situational constraint to infer that the agent's principal motive was to ingratiate himself with his date.

(3) Finally, since the desire to ingratiate oneself is not especially noble, they are not inclined to regard him as morally good when he donates, but are inclined to regard him as morally bad when he steals.

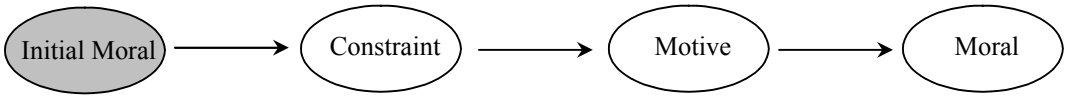
In essence, the idea is that subjects in the two conditions start out by attributing exactly the same level of constraint but that this attribution of constraint leads to very different moral judgments in the different conditions.

This is a very plausible hypothesis, and it might turn out in the end to be correct. Still, it seems to us that recent work in moral cognition is pointing in a very different direction. This work suggests that it is simply a mistake to suppose that people first go through an initial stage in which they are engaged in a purely objective attempt to understand what happened in a given case, followed by a subsequent stage in which they use this information to make moral

judgments. Instead, it seems that the whole process is suffused with moral considerations from the start.

Thus, consider the criteria people ordinarily use to arrive at intuitions about whether an agent *caused* certain outcomes and did so *intentionally*. One might initially suppose that people arrive at these intuitions using some kind of purely objective, entirely non-moral process. But that appears not to be the case. Instead, it seems that people's intuitions about causation and intentional action can actually be influenced by their moral judgments. Although this finding may at first seem counterintuitive, it has emerged in numerous experimental studies, and there is now more than enough evidence to indicate that the effect truly does exist (Alicke 2000; Cushman, Knobe & Sinnott-Armstrong forthcoming; Knobe 2005; Knobe & Fraser 2008; Leslie, Knobe & Cohen 2006; Nadelhoffer 2006; Phelan & Sarkissian 2008; Roxborough & Cumby forthcoming; Young et al. 2006).

Existing research using this new approach has not yet looked specifically at intuitions about freedom and constraint, but we think that there is good reason to suspect that the concept of freedom will turn out to be similar to the various concepts that have already been studied. We therefore hypothesize that the process works more like this:



The key idea behind this hypothesis is that people's moral judgments are actually influencing their intuitions about freedom and constraint. People start out by making a certain kind of 'initial moral judgment' (which we discuss in further detail below). This initial moral judgment then influences people's intuitions about freedom and constraint, which in turn influence a second

kind of moral judgment – namely, the kind of moral judgments about the agent that were examined in Reeder's original experiments.

If this approach is on the right track, we now have at our disposal a new way of explaining Reeder's puzzling results. The puzzle first arose because we assumed that the agent was under exactly the same level of constraint in the two conditions and it therefore seemed strange that people would end up making such different moral judgments about the agent. But there is no reason to suppose that subjects actually conceive of the vignettes in quite that way. The very fact that subjects arrive at different initial moral judgments in the different conditions may lead them to ascribe *different* levels of constraint. People might feel that the agent is somehow more constrained in the morally good condition than in the morally bad one. Indeed, in that latter condition, they may actually conclude that the agent is acting completely freely.

Study 1

Reeder begins his discussion of freedom and constraint by introducing an example from Aristotle. A captain is sailing in his boat when he encounters a violent storm, and he recognizes that the boat will capsize unless he throws a large item overboard. The question now is whether it would be right to say that this captain has been forced by his situation to throw the item overboard or whether it would be more accurate to say that he throws it voluntarily.

As Reeder notes, Aristotle's own view is that cases like this one have an intermediate status but that they partake more of the elements of free action than of action under constraint (*NE* 1110a10-11). It seems to us, however, that people's ordinary intuitions about such cases show a more complex pattern. Specifically, we predict that people's intuitions about whether the

captain acted freely will depend on their judgments as to whether or not it was morally right or wrong to throw the large item overboard.

To test this hypothesis, we conducted a simple experiment.

Method

Subjects were 52 people spending time in a Durham, NC mall. Each subject was randomly assigned either to the *morally neutral* condition or to the *morally bad* condition.

Subjects in the morally neutral condition received the following vignette:

> While sailing on the sea, a large storm came upon a captain and his ship. As the waves began to grow larger, the captain realized that his small vessel was too heavy and the ship would flood if he didn't make it lighter. The only way that the captain could keep the ship from capsizing was to throw his wife's expensive cargo overboard.
>
> Thinking quickly, the captain took her cargo and tossed it into the sea. While the expensive cargo sank to the bottom of the sea, the captain was able to survive the storm and returned home safely.

These subjects then received two questions. The first was taken directly from Reeder and Spores' (1983) original experiment. This question asked subjects to indicate how moral the ship captain was on a scale from 1 ('very immoral') to 7 ('very moral'). A second question then probed subjects' intuitions about the degree to which the agent was constrained. This question asked subjects whether they agreed or disagreed with the sentence:

- The captain was forced to throw his cargo overboard.

Subjects answered this latter question on a scale from 1 ('disagree') to 7 ('agree'), with the midpoint marked 'in between.'

Subjects in the morally bad condition received a vignette that was exactly the same, except that we changed the identity of the large item so as to alter its moral status. This second vignette read as follows (with changes marked in italics):

> While sailing on the sea, a large storm came upon a captain and his ship. As the waves began to grow larger, the captain realized that his small vessel was too heavy, and the ship would flood if he didn't make it lighter. The only way that the captain could keep the ship from capsizing was to throw *his wife* overboard.
>
> Thinking quickly, the captain took *his wife* and tossed *her* into the sea. While the *captain's wife* sank to the bottom of the sea, the captain was able to survive the storm and returned home safely.

Subjects in this second condition received exactly the same questions as those in the first, except that the second question asked whether they agreed or disagreed with the sentence:

- The captain was forced to throw his wife overboard.

The order of questions was counterbalanced here and in all of the other experiments reported in this article, but there were no significant order effects in any of the studies.

Results and Discussion

As expected, subjects in the morally neutral condition rated the captain as more moral (M = 5.3) than did subjects in the morally bad condition (M = 2.4), t (50) = 5.7, $p < .001$.

The real question was whether subjects in these different conditions would have different intuitions about whether the captain was 'forced.' There, we found that subjects actually were more inclined to say that he was forced in the morally neutral condition (M =4.6) than in the morally bad condition (M = 1.9), t (50) = 4.7, $p < .001$.

What we see here is a surprising connection between people's moral judgments and their intuitions about freedom and constraint. Subjects in the different conditions had different intuitions about whether the agent was 'forced' to perform a behavior, but it seems that the one major difference between these conditions lies in the moral status of the behavior the agent performed. So it looks as though people's moral judgments are somehow having an impact on their intuitions about freedom.

Study 2

For our second study, we wanted to extend this basic result to the kinds of cases that show Reeder's asymmetry – cases in which another person pressures the agent to perform either a morally good or a morally bad act. Reeder has already demonstrated that these cases show an intriguing asymmetry in people's moral judgments; we wanted to know whether they would also show an asymmetry in people's intuitions about freedom and constraint.

Method

Subjects were 56 people spending time in a Durham, NC mall. Each subject was randomly assigned either to the *good behavior* or *bad behavior* condition. Subjects in the good behavior condition read the following vignette:

> At a certain hospital, there were very specific rules about the procedures doctors had to follow. The rules said that doctors didn't necessarily have to take the advice of consulting physicians but that they did have to follow the orders of the chief of surgery.
>
> One day, the chief of surgery went to a doctor and said: 'I don't care what you think about how this patient should be treated. I am ordering you to prescribe the drug Accuphine for her.'

> The doctor had always disliked this patient and actually didn't want her to be cured. However, the doctor knew that giving this patient Accuphine would result in an immediate recovery.
>
> Nonetheless, the doctor went ahead and prescribed Accuphine. Just as the doctor knew she would, the patient recovered immediately.

Subjects in the morally bad condition were given a vignette that was almost exactly the same, except that the doctor ends up performing a morally bad behavior:

> At a certain hospital, there were very specific rules about the procedures doctors had to follow. The rules said that doctors didn't necessarily have to take the advice of consulting physicians but that they did have to follow the orders of the chief of surgery.
>
> One day, the chief of surgery went to a doctor and said: 'I don't care what you think about how this patient should be treated. I am ordering you to prescribe the drug Accuphine for her.'
>
> The doctor really liked the patient and wanted her to recover as quickly as possible. However, the doctor knew that giving this patient Accuphine would result in her death.
>
> Nonetheless, the doctor went ahead and prescribed Accuphine. Just as the doctor knew she would, the patient died shortly thereafter.

All subjects then received two questions. As in the previous study, the first of these was Reeder's original question about whether the agent was moral or immoral. The second question then asked subjects whether they agreed or disagreed with the sentence:

- Given the rules of the hospital, the doctor was forced to prescribe Accuphine.

Subjects answered this latter question on a scale from 1 ('disagree') to 7 ('agree'), with the midpoint marked 'in between.'

Results and Discussion

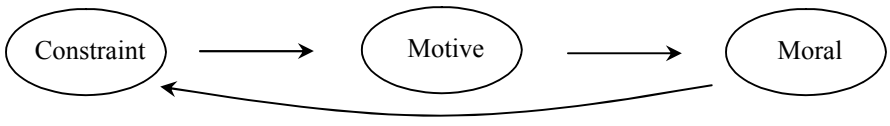
On the question about how moral the doctor was, subjects in the bad action condition said that the doctor was immoral ($M = 2.1$) whereas subjects in the good action condition did not say that he was moral ($M = 3.8$). This simply replicates the results of Reeder's earlier work.

More importantly, we again found an effect of moral judgments on people's intuitions about freedom and constraint. Subjects were significantly more inclined to say that the agent was 'forced' in the good action condition ($M = 4.9$) than in the bad action condition ($M = 2.9$), t (54) = 3.2, $p < .005$.

In other words, it appears that Reeder's asymmetry extends more deeply than one might have expected. Not only do subjects make different moral judgments in the different conditions, they also arrive at different conclusions about the degree to which the agent was free or constrained. What we need now is a specific hypothesis that can explain the surprising connection we find here between people's moral judgments and their intuitions about freedom.

Hypotheses

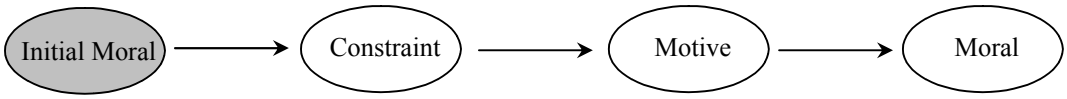
An obvious first hypothesis would be that the asymmetry we find in people's intuitions about freedom is really just a by-product of Reeder's original asymmetry. On this hypothesis, the cognitive process ends up looking more or less like this:



Here the idea would be that people start off by making different moral judgments in the different conditions (Reeder's original asymmetry). Then these moral judgments somehow cause them to go back and adjust their views about whether the agent was acting freely.

Although this hypothesis may turn out in the end to be correct, we think that there is now reason to prefer an alternative approach. We will be developing a hypothesis according to which the asymmetry in people's intuitions about freedom is a genuinely independent phenomenon, which can then serve as part of the explanation for the asymmetry observed in their moral judgments. On this hypothesis, it is not the case that the asymmetry in people's intuitions about whether the doctor was free or constrained is simply a by-product of Reeder's original asymmetry; rather, the asymmetry in people's intuitions about freedom is part of the *explanation* of Reeder's asymmetry.

As we indicated above, our view is that the best way to make sense of this possibility is to posit a role for moral judgment at two distinct steps in the process:



The basic suggestion here is that people are actually making two different *kinds* of moral judgments. They start out by making a certain 'initial moral judgment,' which then affects their intuitions about freedom and constraint, which in turn affects the sort of moral judgment probed in Reeder's original study.

Study 3

Before we can begin testing this hypothesis empirically, we will need to introduce a little bit of additional complexity into our account. In particular, we need to expand our scope and consider a wider range of behaviors.

Thus far, we have been concerned exclusively with the behaviors that the agent actually performed. We now propose to shift the focus over to the behaviors the agent chose *not* to perform. Hence, in the example discussed in Study 2, the aim will be to focus on the option of *not* prescribing Accuphine.

The basic logic of this approach is simple. Fundamentally, it seems that judgments of freedom and constraint are not really judgments about the behavior that the agent actually performs; rather, they are judgments about whether it was open to her to do anything else. Speaking loosely, an agent acts 'freely' when she has other options open, while she is 'constrained' when all of the other options have been closed off. So people's intuitions about whether a given act is free or constrained really depend on their intuitions about whether it was open to the agent to choose any other option.

To capture this additional complexity, we will need to adopt a somewhat more sophisticated account. Instead of supposing that people go directly from an initial moral judgment to an intuition about constraint, we will posit an intermediate step whereby people think about whether the agent had any other options open. The first stages of the process might then go like this:

On this hypothesis, people's initial moral judgments directly affect their intuitions about whether it was open to the agent to *not* prescribe Accuphine. This first intuition then affects their intuitions as to whether the agent's actual behavior of prescribing Accuphine was free or constrained.

With this conceptual background in place, we can now test two key claims of our model: (1) that people's intuitions about the other options can be affected by some kind of moral judgment and (2) that these intuitions are not simply by-products of the specific type of moral judgment that was measured in Reeder's original studies.

Method

Subjects were 60 people spending time in a Durham, NC mall. The experimental design was exactly the same as that in Study 2, except that we replaced the question about whether the agent was 'forced' to do what he actually did with a question that asked explicitly about the behavior that the agent chose not to perform. This new question asked subjects whether they agreed or disagreed with the sentence:

- Given the rules of the hospital, the doctor did not really have the option of not prescribing Accuphine.

Subjects rated this sentence on a scale from 1 ('disagree') to 7 ('agree').

Results

Once again, we replicated Reeder's original finding that subjects regard the agent as immoral in the bad behavior condition ($M = 2.4$) but as not particularly moral in the good behavior condition ($M = 4.5$).

The important results, however, concerned people's intuitions about the behavior that the agent chose not to perform. There, we found a highly significant effect such that subjects were more inclined to think that the agent 'did not really have the option' of performing this behavior in the good behavior condition ($M = 5.5$) than they were in the bad behavior condition ($M = 2.9$), $t\,(58) = 4.5$, $p < .001$.

We then conducted a mediational analysis to examine the relationships among the different variables in the study. This analysis indicated that people's intuitions about whether the agent had another option mediated the effect of the difference between conditions on their moral judgments about the agent.[1] In other words, part of the reason why people made different moral judgments about the agent in the different conditions was that they had different views about whether or not the agent had any other option.

Discussion

This last study yielded two major results. First, it seems that some kind of moral judgment is having an impact on people's intuitions about the behavior the agent chose not to perform. After all, the two conditions lead to two very different intuitions, and yet it looks like the only major difference between these two conditions lies in the moral properties of the behaviors described.

But, second, when we look at the particular type of moral judgment that was actually measured within the study, we find that people's intuitions are not simply by-products of that specific type of moral judgment. Indeed, the causal chain appears to go in exactly the opposite

[1] As we noted above, the difference between conditions had a significant effect both on 'option' judgments and on moral judgments. The option judgments and the moral judgments were themselves significantly correlated ($r = .67$, $p < .001$). To test for mediation, we ran a series of regression analyses. When condition and option were entered simultaneously, the regression coefficient measuring the relationship between condition and option went from $\beta = .46$, $p < .001$ to $\beta = .16$, $p > .15$. A Sobel test showed that this reduction was significant, $Z = 3.76$, $p < .001$.

direction. The difference between the two conditions is affecting people's intuitions about the behavior the agent chose not to perform, which is in turn affecting the moral judgments measured in our study.

It seems to us that the best way to make sense of this pattern of results is to posit an 'initial moral judgment' that differs in some way from the moral judgment that was actually measured within the study. We can then propose a more complex causal chain:

On this model, people's initial moral judgments affect their intuitions about whether any other options are open, which affect their intuitions about whether the agent acted freely, which eventually affect the kind of moral judgments measured in our study.

The trick now is to develop a picture of people's ordinary understanding that enables us to make sense of this sort of effect – a picture that allows us to see how people's moral judgments about an option might impact their intuitions about whether an agent is free to perform it.

General Discussion

We have criticized certain aspects of Reeder's specific theory, but we are very much in sympathy with his basic approach. Reeder's principal claim is that one will never be able to arrive at a proper understanding of people's ordinary conceptions of freedom and constraint if one continues working within the framework of classic attribution theory. Instead, one needs to return to the problem with fresh eyes and try to think about the role that the ordinary notions of freedom and constraint really play in people's lives.

We think that this general approach is right on target. However, we are not quite in agreement with Reeder's specific claims about what the role of people's concept of freedom actually is. Reeder offers a picture in which this concept serves as a tool for making accurate inferences about people's mental states. In our view, this sort of picture cannot capture the full richness of people's ordinary understanding. It may well be that social scientists sometimes use the concept of freedom as a tool for accurately inferring mental states, but the evidence suggests that people's ordinary understanding of freedom is quite a bit more complex. People's ordinary understanding does not appear to form a part of some kind of objective, impartial attempt to understand human behavior. On the contrary: it appears that people's ordinary understanding of freedom is wrapped up in a very fundamental way with *moral* considerations.

In the space remaining, we will be offering a specific proposal about how this process might proceed. However, we want to begin by emphasizing that the principal aim of our paper is not so much to advance this one specific proposal as to argue for the more general claim that people's intuitions about freedom are somehow being influenced by an initial moral judgment.

Deriving the asymmetry in intuitions about freedom

On the proposal we wish to advance, people think that an agent acts freely to the extent that certain other options are 'open,' and they think that an agent acts under constraint to the extent that all other options are 'closed.' The key idea then is that people's conceptions of openness and closedness are not at all the sorts of things that could form the basis of an entirely impartial scientific theory. Instead, these conceptions are connected on a very basic level to certain moral judgments. In particular, one of the factors that can make an option seem more

'closed' is that it is regarded as morally wrong, and one factor that can make an option seem more 'open' is that it is regarded as morally right.

It now becomes possible to reach a better understanding of our proposed 'initial moral judgment' and how that judgment differs from the judgments measured in Reeder's original studies. The key thing to notice is that the initial moral judgment is not actually a judgment about the behavior the agent performed. Rather, it is a judgment about a behavior the agent chose *not* to perform. So the basic approach will be to argue that people's moral judgments about the various other options are impacting their intuitions about whether these options are open or closed.

To illustrate this approach, we can return to the case of the captain who throws his wife's cargo overboard. Here, people feel that the option of *not* throwing the cargo overboard is fundamentally closed. But this conclusion does not follow straightforwardly from some kind of purely impartial scientific inquiry. Instead, it is the product of a value judgment people make about this other option. People feel that it is far more important to stop the ship from capsizing than it is to save the cargo. For this reason, they conclude that the option of not throwing the cargo overboard is so bad that it is not even worth considering, and they end up regarding this option as 'closed' by the situation. The captain appears to have been forced to act as he did.

Now suppose we turn to the case in which the captain throws his wife overboard. People arrive at very different intuitions about this case because they start out with very different value judgments. They feel that there is something deeply morally right about the option of not throwing one's wife into the sea, and for that reason, they conclude that the option of not throwing the wife overboard is fundamentally open. They therefore reject the idea that the captain was forced to do what he did.

In this way, people's moral judgments can affect their intuitions about whether an option is open or closed and, thereby, about whether the agent is free or constrained.

Deriving Reeder's asymmetry

With these ideas in the background, we can now propose a new and radically different explanation for Reeder's asymmetry. On our explanation, subjects in the bad action condition go through a process that looks roughly like this:

(1) Since the behavior itself is morally wrong, the option of *not* performing the behavior is morally right. (*Initial moral judgment*)

(2) Since the option of not performing the behavior is morally right, this option is fundamentally open.

(3) Since the agent had another option open, the action she actually performed was done freely.

(4) Since the action she actually performed was done freely, it has real significance for our moral judgments about her character. (*Reeder's original asymmetry*)

The important thing to notice about this explanation is the role it assigns to people's moral judgments. On our hypothesis, moral judgments do not simply appear as an extra step added on at the end. Instead, the whole process is suffused with moral considerations from the very beginning.

References

Alicke, M.D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin, 126,* 556-574.

Aristotle. (1985). *Nicomachean ethics*. (T. Irwin, Trans.). Indianapolis, Ind: Hackett Pub.

Cushman, F., Knobe, J. & Sinnott-Armstrong, W. (2008). Moral appraisals affect doing/allowing judgments. *Cognition, 108*, 281-289.

Knobe, J. (2005). Theory of mind and moral cognition: Exploring the connections. *Trends in Cognitive Sciences*, 9, 357-359.

Knobe, J. & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. In W. Sinnott-Armstrong, *Moral Psychology*, Cambridge, MA: MIT Press. 441-448.

Leslie, A., Knobe, J. & Cohen, A. (2006). Acting intentionally and the side-effect effect: 'Theory of mind' and moral judgment. *Psychological Science, 17*, 421-427.

McGraw, K. M. (1985). Subjective probabilities and moral judgments. *Journal of Experimental Social Psychology, 21*, 501-518.

McGraw, K. M. (1987). Outcome valence and base rates: The effects on moral judgments. *Social Cognition, 5*, 58-75.

Nadelhoffer, T. (2006). On trying to save the simple view. *Mind & Language, 21:5,* 565-586.

Phelan, M. & Sarkissian, H. (2008). The folk strike back; or, Why you didn't do it intentionally, though it was bad and you knew it. *Philosophical Studies, 138,* 291-8.

Roxborough C. & Cumby, J. (forthcoming). Folk psychological concepts: Causation. *Philosophical Psychology*.

Trafimow, D., & Trafimow, S. (1999). Mapping perfect and imperfect duties onto hierarchically and partially restrictive trait dimensions. *Personality and Social Psychology Bulletin, 25,* 686-695.

Vonk, R., & van Knippenberg, A. (1994). The sovereignty of negative inferences: Suspicion of ulterior motive does not reduce the negativity effect. *Social Cognition, 12*, 169-186.

Young, L., Cushman, F., Adolphs, R., Tranel D., and Hauser, M. (2006). Does emotion mediate the relationship between an action's moral status and its intentional status? Neuropsychological evidence. *Journal of Cognition and Culture, 6(1-2)*, 265-278.