Directions and Challenges in Studying Folk Concepts and Folk Judgments

Bertram F. Malle* & Steve Guglielmo**

We would like to discuss two topics, one specific, the other more general. We first track an emerging trend in current research on the relation between morality and judgments of intentionality. Then we take up the general question of how we can study such folk concepts as *intentionality*.

From badness to broader hypotheses

The multitude of research examining folk judgments of intentionality and morality following Joshua Knobe's (2003a) intriguing findings has largely sought to explain the disparity in intentionality judgments for immoral vs. nonmoral (neutral) actions. Knobe discovered that people do not regard the presence of skill (or even intention) as necessary for judging an action intentional when the action is morally bad. The essence of the question is *why*, when considering a morally charged action, people seem to abandon their basic concept of intentionality (Malle & Knobe, 1997) – in which, it had appeared, desire, belief, intention, awareness, and skill must each be present to consider an action intentional. To answer this question, many researchers have focused on the role of blame.

* Institute of Cognitive and Decision Sciences and Department of Psychology, University of Oregon. Email: bfmalle@darkwing.uoregon.edu.

** Institute of Cognitive and Decision Sciences and Department of Psychology, University of Oregon.

Whether intentionality judgments are used to inform judgments of blame (Knobe & Mendlow, 2004) or vice versa (Nadelhoffer, forthcoming), a substantial portion of recent research has aimed at elucidating the interplay between intentionality and blame. A number of recent studies have made interesting and theoretically-relevant adaptations to Knobe's original vignettes, manipulating, for example, the general moral valence of the outcome, the agent's degree of skill, and the scope of the relevant intentions. The underlying assumption has been that blame and intentionality judgments are linked and that the discrepancy in intentionality judgments for immoral and nonmoral actions can be explained by appealing to this relationship.

Young, Cushman, Adolphs, Tranel, and Hauser (this issue) provide intriguing new results that challenge this relationship (while also nicely illustrating the interdisciplinary potential of the phenomena of interest here). The researchers find that in patients who display severely compromised emotion processing, Knobe's (2003a) original discrepancy is fully replicated. If we assume that the moral sentiment of blame has a strong component of negative affect, and if we assume that the patients truly had no emotional response to Knobe's vignette, the finding suggests that an affective process of blaming is not necessary for the discrepant intentionality judgment effect to obtain.[1] This result adds to the growing suspicion that Knobe's findings (and the many replications) are not, or not exclusively, an issue of moral sentiments but at least partially a result of conversational demands, conceptual vagueness, and methodological problems (Adams, this issue; Adams & Steadman, 2004a, 2004b; Malle, this issue). This would not make the findings uninteresting; we merely have to refrain from drawing strong conclusions from them – for example, that *intentionally* has nothing to do with *intend*, or that *intentionally* has two distinct literal meanings (Knobe & Burra, this issue).

What has been overlooked in most of the recent theoretical and empirical investigations is Knobe's (2003a) equally intriguing finding that

[1] The findings are also compatible with the intriguing possibility that moral emotions are not emotions – that they are not processed the same way, and in the same brain networks, as emotions and nonmoral evaluations.

the folk also are willing to ascribe intentionality to a morally good outcome (even when the agent lacks skill). Of course, such a phenomenon cannot be accounted for by appealing to the role of blame alone. Nadelhoffer (2005) suggests that sufficiently blameworthy or praiseworthy actions override any skill requirement and are sufficient for making judgments of intentionality. Knobe and Burra (this issue) are careful to state that the function of intentionality judgments (not intention judgments) may just be to praise and blame. So perhaps intentionality judgments operate leniently at both moral extremes – that is, if an agent lacks skill but his action is sufficiently morally extreme, people will deem the action intentional. To withhold praise for a positive outcome (by saying "she didn't do it intentionally") may then be just as aversive as to withhold blame for a very negative outcome.

A concept that is relatively symmetrical with respect to positive and negative outcomes is responsibility – the linking of an agent to an outcome that in turn *gives rise to* blame and praise (Weiner, 1995). Are then judgments of intentionality, when used in morally extreme situations, simply assimilated to an ascription of responsibility? Responsibility judgments are indeed closely related to intentionality judgments (Weiner, 1995; Malle, Moses, & Baldwin, 2001). Obviously, a judgment that somebody intentionally brought about a (negative) outcome X is normally sufficient for a responsibility ascription. But even when X was brought about unintentionally, responsibility can be ascribed – if there was an obligation to avoid X and the agent (aware of X) intentionally failed to prevent X. The results in Knobe's side-effects scenario (the CEO is seen as "intentionally" harming the environment without intending it) may thus hold because the agent in the scenario refuses to take protective action, thereby violating his responsibility to protect the environment. And because only an intentionality scale, not a responsibility scale, is offered, people say he did it "intentionally" while perhaps meaning to say he should be held responsible for the harm.

If this reasoning is correct, several new predictions follow. First, we should see lower rates of intentionality judgments among participants who care very little about the moral outcome (e.g., the harmed environment), because they don't ascribe an obligation to the agent to avert the harm. Second, as Alfred Mele (this issue) suggests, we should see lower intentionality judgments if the CEO declares he is well aware of

this obligation, feels bad about not meeting it, but feels even more obligated to further the health of his company and his employees. We are currently testing people's intentionality judgments in response to such a scenario.

Professor Mele offers another important point that may strengthen an "emotion-free" (or at least blame-free) account of the puzzling findings on intentionality. He suggests that an agent must intend to do *something* in order for other things she does to be intentional. In the absence of any intention, even great damage (e.g., a car accident that kills a family of 5 because the driver fell asleep at the wheel) will not be seen as having been brought about intentionally. But how close (in time or content) must a behavior (or side-effect) in question be to fall under the scope of the "primary" (and unquestionably intended) action and be considered intentional as well? Pulling the trigger obviously suffices for people to believe that the agent *killed* the target/victim even if the execution of the relevant behavior was marred by imperfections. But would raising one's rifle suffice, with a shot going off before the agent even pulls the trigger? What if the person decided to delay the killing and then, just as she forms that decisions, a shot goes off?

It is evident that knowledge of an agent's specific mental states prior to and during an action influences both intentionality judgments about that action and evaluative judgments of blame and praise. However, the role of post-action mental states may need to be investigated as well because they, too, may play a role in guiding folk judgments of blame, praise, and intentionality. We are currently exploring situations in which an agent feels varying degrees of remorse following a negative outcome (or pride following a positive outcome). Finding out that an agent regrets her immoral behavior may tone down ratings of blame. But what will happen to intentionality judgments? If they stay at the same high level, the blame hypothesis is weakened, especially if responsibility judgments also stay at a high level. If intentionality judgments go down with blame judgment, the blame hypothesis regains strength. No theoretical model currently predicts whether declarations of pride will moderate praise judgments and/or intentionality judgments, so our results should help inform these models.

Do we know folk concepts when we see them?

We now turn to our second topic: what we can know about folk concepts. Our discussion is inspired by Anna Wierzbicka's (this issue) commentary and is thus partially a reply to her.

Professor Wierzbicka doesn't have much good to say about the research endeavors in the present issue. In particular, she expresses concerns about what we can know about the concept of intentionality – its meaning, whether it is a folk concept, and whether it is a universal concept. She writes, "no ordinary folks have a clue as to what the word *intentionality* means and they don't have it in their vocabulary." And: "I do not know exactly what the author has in mind when he talks about "intentionality."

Apparently, Professor Wierzbicka has not had an opportunity to read the article that underlies many of the discussions here, in which Joshua Knobe and one of us offered a series of empirical studies of people's concept of intentionality (Malle & Knobe, 1997). In these studies, we didn't ask our participants what "intentionality" means. This is indeed a technical term. Instead, we asked participants *what it means when somebody does something intentionally* – which is just the definition of the term *intentionality* we had adopted in our paper: the attribute of an action to have been performed intentionally. In people's "folk definitions," the component concepts of *belief, desire, intention*, and *awareness* emerged (whether or not we had provided them with a definition of the term *intentionally* to start with). Unsurprisingly, people's answers only sometimes included the specific words that we as researchers use to refer to these component concepts (e.g., *desire, intention*), but we employed a coding procedure that reliably grouped the various linguistic expressions under these concepts.[2] Hence emerged a first sketch of a model of people's concept of "intentionally-performing-an-action" (in short, their concept of *intentionality*). Additional experimental data supported this model and suggested

[2] Reliability (inter-judge agreement) of a coding procedure does not guarantee validity (i.e., the coding categories group the words at their true boundaries). However, the theoretical model that was derived from the coding was consistent with much previous research, and it was put up for falsification in the article's subsequent experiments.

that people's judgments of intentionality were sensitive not only to *belief*, *desire*, *intention*, and *awareness*, but also to a fifth component, namely skill (Malle & Knobe, 1997).

Our evidence converges to some extent with Professor Wierzbicka's universal semantic analysis, in which concepts such as WANT, THINK, and KNOW are claimed to be universal. However, what this semantic analysis doesn't tell us, and our approach at least strives to uncover, is the relationship among these concepts. For example, the list of conceptual primitives Professor Wierzbicka proposes (e.g., Wierzbicka, 1996) doesn't clarify the distinction between *intending* to do something and *doing it intentionally* (Malle & Knobe, 1997), or the distinction between *wanting* and *intending* (Malle & Knobe, 2001). Our empirical data suggest that people – admittedly, Americans – make these distinctions quite naturally. Supporting the uniqueness of the *intention* concept, Joan Bybee (1997) has done cross-linguistic work showing that, across many languages, the concept of *intending* emerges from the future tense. This finding does not appear to extend to the concept of *goal* or *desire*.

Now, Professor Wierzbicka is free to translate the results of our experimental evidence on conceptual relations into her meta-language, but the meta-language itself doesn't produce this sort of evidence.[3] At the same time, Professor Wierzbicka insists that only the meta-language will allow us to pinpoint the real concepts people have. "Describing human cognition in English (or quasi-English) words like *intentionality*, *agency* and *morality*, in itself imposes the researchers' terms on other people's ways of thinking." This is a misunderstanding. The whole endeavor here is to infer from people's talking (in their own words), acting, choosing, reacting, and protesting what concepts they rely on and what meaning, what semantic components, these concepts have. If researchers then

[3] As a side note, the translation into and from this meta-language is a precarious topic, and there is reason to be skeptical about the intersubjective agreement of such translation. As far as we know, Wierzbicka typically uses her vast knowledge of languages and a dictionary to establish the translational equivalences, not any systematic study of relevant populations. We have tried to understand some of the translations that Professor Wierzbicka offers in her commentary, and we have failed. Certainly, we would not want to present these statements to our participants in future experiments, as Professor Wierzbicka recommended.

use abstract noun forms to label these complex concepts, they are actually better secured against the possibility of merely describing English, Polish, or Mandarin verbs, adjectives, suffixes, and so on. But even when using these semi-technical terms, we have to listen very carefully to what people say, how they say it, and under what circumstances, so that we really are looking at all the data necessary to draw informed scientific inferences. In the end, paying attention to language (which we deem essential in this line of work) and adopting a "meta-language" may actually be incompatible. The very culture-specific ways of expressing and negotiating the concepts people have may reveal interesting psychological aspects of those concepts. Eventually, the hypotheses gleaned within one culture/language will have to be tested in other cultures/languages – but in the specific terms that the community of speakers uses, not in the terms of a meta-language.

Some of the disagreement over the status of folk concepts (universal or not) stems from the fact that Professor Wierzbicka assumes a tight connection between concepts and their corresponding words across languages. Rejecting talk about a folk concept of intentionality, she points out that "very few languages other than English have a word which could be matched in meaning with the technical English word *intentionality*." Similarly, she states that "the word *agency* does not stand for a folk concept because it is a technical, philosophical term which is not part of everyday language." But this is not a compelling argument, for something can be a technical term and still capture a folk concept. As the legal world amply illustrates, people can have a folk concept while "experts" use the same word in their own technical ways (Malle & Nelson, 2003). Moreover, people can have a folk concept without having a word for it (e.g., the concept of *causal history of reasons* plays an important role in people's explanations of behavior but does not correspond to an everyday term; Malle, 1999, 2004). But all this is just prelude to the central question: How do we decide whether something is a (folk) concept? Our position is that only convergent empirical data can make such a case. Consider this evidence about the family of concepts *agency, intentionality*, and *goal-directedness*.

Infant studies suggest that 6-month old babies may have an emerging concept of agency because they respond in meaningful ways to goal-directed actions (Woodward, Sommerville, & Guajardo, 2001). For example,

if a human hand repeatedly grasps one of two objects but then the objects switch position, infants are more surprised if the identical movement occurs (but grasping the "unwanted" object) than if a new movement occurs (grasping the "wanted" object in the changed location). By 14 months, toddlers are sensitive to boundaries between successive intentional actions (Baldwin, Baird, Saylor, & Clark, 2001), and between 14 and 18 months, they learn to distinguish between intentional and accidental behaviors (e.g., they imitate only the former, not the latter; Carpenter, Akhtar, & Tomasello, 1998). By 18 months, children can observe the beginnings of another person's action (on a never-before seen object), infer what the person's goal is, and complete the action themselves even though they have never seen the complete act on this novel object (Meltzoff, 1995). None of these results require verbal responses, yet they all are evidence for (emerging) *concepts*. And none of these results are in themselves conclusive about what the concepts achieve and how they do so. But taken together, they paint a picture of a developing network of sensitivities, distinctions, and categories that eventually maps onto the familiar concepts of (not words) *intentionality, belief, desire*, and so on.

In identifying concepts, folk concepts, and especially universal folk concepts, it is not sufficient to count whether and how many languages have a word "for" this concept (who decides, by the way, whether a given word is a word for *this* concept?). Observational and experimental designs, cognitive and behavioral as well verbal and nonverbal data, within-culture and across-culture samples, person-centered and neuroscientific approaches – all these have to come together to make the strongest possible case for claiming that something is a concept, a folk concept, a universal concept.

REFERENCES

Adams, F., & Steadman, A.
2004a Intentional action in ordinary language: Core concept or pragmatic understanding? *Analysis*, *64*, 173-181.

Adams, F., & Steadman, A.
2004b Intentional action and moral considerations: Still pragmatic, *Analysis*, 64, 268-276.

BALDWIN, D. A., BAIRD, J. A., SAYLOR, M. M., & CLARK, M. A.
2001 Infants parse dynamic action. *Child Development, 72*, 708-717.
C, M., AKHTAR, N., & TOMASELLO, M.
1998 Fourteen- through 18-month-old infants differentially imitate intentional and accidental actions. *Infant Behavior and Development, 21,* 315-330.
KNOBE, J.
2003a Intentional Action in Folk Psychology: An Experimental Investigation, *Philosophical Psychology*, 16, 309–324.
2003b Intentional Action and Side Effects in Ordinary Language, *Analysis, 63*, 190-193.
MALLE, B. F., & KNOBE, J.
1997 The folk concept of intentionality. *Journal of Experimental Social Psychology, 33*, 101-121.
MALLE, B. F., & KNOBE, J.
2001 The distinction between desire and intention: A folk-conceptual analysis. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 45-67). Cambridge, MA: MIT Press.
MALLE, B. F., & NELSON, S. E.
2003 Judging *mens rea*: The tension between folk concepts and legal concepts of intentionality. *Behavioral Sciences and the Law, 21*, 1-18.
MALLE, B. F., MOSES, L. J., & BALDWIN, D. A.
2001b The significance of intentionality. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 1-24). Cambridge, MA: MIT Press.
MELTZOFF, A. N.
1995 Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology, 31*, 838-850.
WEINER, B.
1995 *Judgments of responsibility: A foundation for a theory of social conduct.* New York: Guilford.
WIERZBICKA, A.
1996 *Semantics: Primes and universals.* New York: Oxford University Press.
WOODWARD, A. L., SOMMERVILLE, J. A., & GUAJARDO, J. J.
2001 How infants make sense of intentional action. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 149-170). Cambridge, MA: MIT Press.