On Folk Conceptions of Mind, Agency and Morality*

Paulo Sousa**

Folk conceptions of mind comprise, among other things, concepts of dispositional traits like generosity, stubbornness and wickedness, concepts of propositional attitudes like beliefs, desires and intentions, concepts of emotions like anger, guilt and love, and concepts of mental actions such as choices and decisions. In addition to concepts of mental actions, folk conceptions of agency comprise concepts of behavioral actions, and classifications of types of both behavioral and mental actions – intentional versus unintentional action, free versus non-free action, and free choice versus non-free choice. Folk conceptions of morality comprise, among other things, concepts of norms and obligations, concepts of responsibility, such as blameworthiness or praiseworthiness, and concepts of vices and virtues.[1]

There are many apparent connections between these folk conceptions. The classification of different types of observable behaviors may depend on their mental etiology. A basic notion of action may be part of the folk concept of obligation, if one considers such a concept as a one-place predicate with a placeholder for action (e.g., Jackendoff, 1999).

---

* Thanks to Tom Lawson for criticisms and suggestions.

** Institute of Cognition and Culture, Queen's University Belfast. Email: p.sousa@qub.ac.uk.

[1] Of these three general labels, the one used most expansively here is 'morality.' One could say that there are norms and obligations that are not moral, that there is blameworthiness and praiseworthiness that is not moral, and that there are vices and virtues that are not moral. So, at first glance, folk conceptions of morality are related to something more specific in nature. However, what this more specific nature consists of is an open question that is part of the very topic of this special issue (See Stephen Stich's commentary, this issue, for a related discussion.).

The folk concept of obligation implies some notion of *can*, which in turn implies some concept of free action and free will. If one is responsible (more specifically, is to blame) for some event, there is often the normative expectation that one ought to have avoided bringing about the event or letting it happen. One is considered to be *more* responsible for some event if one is interpreted as acting intentionally and freely. Attributions of responsibility may be driven by an interest in identifying a vicious or virtuous character. Some vices and virtues are mental dispositional traits. Finally, there is the unifying concept of an agent – the person who has some dispositional traits, who has beliefs, desires and obligations, who makes choices, who forms intentions and acts, and who is deemed responsible for her actions and the consequences of her actions.

The main goal of this special issue is to discuss these folk conceptions and the nature of their connections, as well as the architecture of the mind that frames these conceptions and connections.

Of course, many aspects of this general topic have been under scrutiny in the psychological and socio-cultural sciences for decades and in philosophy for centuries, and it will be important to explicate the primary concern of the inquiry here. Shaun Nichols (this volume) makes a distinction between three types of approaches to the study of the folk concepts of free will and moral responsibility: a descriptive one, in which the aim is simply to characterize these folk concepts; a substantive one, in which the aim is to determine how the world is and whether folk views are correct given how the world is; and a prescriptive one, in which the question is whether these folk concepts and the practices that presuppose them should be revised. In a pertinent way, one can extend Nichols' distinctions to other folk concepts that are the focus of this special issue. Thus, descriptively, one may be interested in characterizing the folk concepts of intentional action and intention. Substantively, one may be interested in which type of action an intentional action is in itself and in which type of mental state an intention is in itself, and whether the folk concepts of intentional action and intention are correct. Finally, prescriptively, one may, *à la* Paul Churchland, propose that these concepts and their implied practices be completely revised.[2] Although

[2] In alluding to Paul Churchland's eliminative materialism here (see, e.g., Churchland, 1979), I am envisaging some similarities between Nichols' discussion and debates about reduction and elimination in the philosophy of science (see McCauley, 1986). However,

in this special issue one will find some discussions regarding substantive and prescriptive questions, the volume's primary concern is with the descriptive one, and this introduction follows suit.

The psychological sciences are the ones that have most consistently pursued a descriptive approach to folk concepts with experimental methods. However, in relation to the topic of this issue, the flow of research has led the theoretical landscape towards compartmentalization. Thus, one finds a "Theory of Mind" literature that deals mainly with folk conceptions of mind and some aspects of folk conceptions of action apart from the literature on moral reasoning (see, e.g., Leslie, 1994; Nichols & Stich, 2003; Wellman, 1990). Moreover, one finds some kind of broad division in the literature on moral reasoning between those who concentrate on folk conceptions of norms (Turiel, 1983, 1998) and those who deal with folk attributions of responsibility and blame (see, e.g., Alicke, 2000; Shaver, 1985; Weiner, 1995).[3] More recently, though, there has been an increasing interest not only in making bridges across these literatures, but also in introducing into the discussion new concepts that foster the bridging (see, e.g., Bloom, 2004; Harris, 2002; Hauser, *forthcoming*; Kadish, 2005; Knobe, 2003a, 2003b; Malle & Nelson, 2003; Malle et al., 2001; Morton, 2002; Nadelhoffer, 2005; Nahmias et al., 2005, Nichols 2004a, 2004b; Nunes & Harris, 1998; Turner & Nahmias, *forthcoming*; Wellman & Miller, *forthcoming*; Woolfolk et al., *forthcoming*).[4]

---

to be more precise, it is worth noticing that elimination in the epistemological sense is rather a substantive question and does not imply that folk views could or should be *psychologically* eliminated – after all, folk psychology may be epistemologically eliminated by a well-developed neuroscience without it being possible or desirable that humans ordinarily think about other humans without using folk psychological concepts (cf. folk physics). Churchland tends to conflate epistemological and psychological elimination because he holds a very strong hypothesis about the plasticity of the mind and is too optimistic about education.

[3] For an interesting discussion of both moral literatures, see Darley & Shultz 1990.

[4] Analytic philosophers have for a long while pursued a descriptive (even if, often, not a purely descriptive) approach in their a priori conceptual analyses (see Nichols, this volume). However, more recently, with the burgeoning field of experimental philosophy, analytic philosophers have started to approach the analysis of folk concepts with empirical methods, and their interests are to a great extent similar to those of psychologists and cognitive scientists. Drawing from a tradition that has already attempted to analyze the concepts and connections mentioned initially, it is not a coincidence that experimental philosophers play an important role in this recent literature and in this volume.

Another goal of this special issue is to create a forum for this more recent literature. The bulk of the volume is comprised of three parts. Target articles propose hypotheses and/or report empirical evidence concerning different aspects of the topic. Commentaries envisage alternative hypotheses and/or present additional evidence related to points raised by the articles. In the last part, some of the target articles' authors respond to the commentaries.

In the remainder of this introduction, I will go through the articles and commentaries, highlighting some of the hypotheses, arguments, and evidence discussed by the contributors. I will sort my description into three thematic chunks: folk concepts of norms and mental states, folk concepts of free will and moral responsibility, and the folk concept of intentional action and its relation to moral judgments. However, before dealing with these themes, let me point out two general questions that one should keep in mind in reading this volume. First, whose conceptions are at stake here? Who are the folk? American undergradutes, undergraduates in general, Caucasian Americans, Americans, western people, Indians, non-western people, all normal humans? This first type of question, raised by many of the contributors, is fundamental in evaluating the import of the hypotheses envisaged throughout this volume. Another related type of question is this: to what extent are descriptions of differences and similarities in folk conceptions ethnocentric? To see the relevance of this question, let me indicate one of the arguments raised by Anna Wierzbicka (this volume) in this respect. Wierzbicka argues that the expression "folk conceptions of mind" is somewhat paradoxical because it carries some implication of universality when in fact *mind* in itself is a culturally-bound concept:

> (. . .) *mind* is an English word with no exact semantic equivalent in other languages, or indeed in older English (. . .) 'Mind' is an important folk concept in modern English, just as 'duša' is an important folk concept in modern Russian, 'kokoro' in modern Japanese, 'maum' in modern Korean, and so on (. . .) All these words imply a dichotomous model of a person, in which a person has two main parts: a visible one (the body), and an invisible one. Linguistic evidence shows that the first, visible part is conceived of in essentially the same way in all cultures, as all languages have a word corresponding in meaning to the English word *body* (in the relevant sense). The other main part of the person, however, is conceived of differently in

> different cultures. Roughly speaking, the (contemporary) English word *mind* presents this part as primarily an organ of thinking and knowing, whereas the Russian word *duša*, the Japanese word *kokoro* and the Korean word *maum* link it in various ways with feeling, wanting, and choosing between what is 'good' and what is 'bad'. (Wierzbicka, this volume, p. 165)

Norm Concepts & Psychological Concepts

In their article "Developing Conceptions of Responsive Intentional Agents," Henry Wellman and Joan Miller argue that a universal feature of folk psychology[5] is the concept of a *responsive intentional agent.*

The agent is construed as intentional because she is supposed to have mental states such as beliefs, desires and intentions and is supposed to act according to her mental states, that is, intentionally. The agent is construed as responsive because she is supposed to be sensitive to the circumstances that constrain her actions – and especially to her social circumstances, the social norms that constrain her actions, since "it is this social responsiveness that is central to folk psychology."

They also argue that this concept of a responsive intentional agent provides a more complete characterization of the folk conception, which is missing in the current literature:

> . . . contemporary discussions of theory of mind, with their emphasis on the mental states of individual agents, tend to portray persons as autonomous agents – intentional actors whose actions are determined by individual choices, preferences and beliefs. This is important, but only part of the story because persons are, equally, responsive agents – intentional actors whose actions are influenced by social norms . . . A notion of responsive intentional agency helps capture this more balanced everyday conception . . . (Wellman & Miller, this volume, p. 28)

[5] They use the expression 'folk psychology' as an umbrella term that covers many of the folk concepts I mentioned initially – in particular, it covers both concepts of norms and concepts of psychological states.

According to Wellman and Miller, some core features of this agent concept are already shown in infant cognition, and the development of these features during childhood leads to an understanding of persons that is arguably universal. Three important aspects of this childhood development should be emphasized, the authors claim:

First, the two sides of the agent concept tend to go hand in hand in development. Thus, in holding someone accountable for a moral violation, children show an understanding of the difference between intentional and unintentional actions and an understanding that a social norm applies to the action; they also show an understanding of the type of social norm involved – moral instead of conventional. Secondly, there are fundamental conceptual changes in childhood. For instance, children develop a fully-fledged representational concept of belief from a previous non-representational one, which is indicated by the fact that at a certain point all normal children pass a range of false-belief tasks. Finally, conceptual revolutions are not immune to cultural-communicative influences – meta-analyses of false-belief tasks in different societies and results of false-belief tasks with deaf children show significant differences in developmental timing.

Wellman and Miller also argue that, despite this overall childhood convergence, afterwards, cultural experiences "can lead to adult endpoints of folk psychology that are in many respects strikingly diverse across cultures and languages." As an example of a striking difference in folk psychologies, they indicate the divergent views of what is normative (the domain of social norms) and of what is discretionary (the domain of personal choice) held by western individualist societies (more specifically, North American societies) and by non-western collectivist societies (more specifically, Indian societies). In order to explicate that, let me introduce two different but related sets of concepts.

Not to put too fine a point on it, the three basic *deontic* concepts are inter-defined as follows:

(i) What is obligatory: *what is at the same time permissible to do and prohibited not to do*;
(ii) What is prohibited: *what is not permissible to do*;
(iii) What is permissible: *what is obligatory to do* or *what is neither obligatory nor prohibited to do*.

For the sake of illustration, consider the action of helping someone as part of what is obligatory, the action of killing someone as part of what is prohibited, and the actions of helping someone (what is obligatory) and of eating ice cream (what is neither obligatory nor prohibited) as part of what is permissible.

The second set contains just a general opposition between what is normative and what is discretionary:

(i) What is normative: *what is obligatory to do* or *what is prohibited to do*;
(ii) What is discretionary: *what is neither obligatory nor prohibited to do*.

Here, the actions of helping someone and of killing someone are both under the scope of what is normative; the action of eating ice cream is simply part of what is discretionary.[6]

This general concept of what is normative entails a sense of what is right and a sense of what is wrong. In other words, what is obligatory, what one ought to do, is the right thing to do, and what is prohibited, what one ought not to do, is the wrong thing to do. On the other hand, what is discretionary is what is neither the right thing to do nor the wrong thing to do – it is simply what is "all right" to do, a question of personal preference-based choice.

Wellman and Miller claim that there are some important differences between American and Indian views of what is normative and what is discretionary.[7]

Consider actions related to friendship, loyalty and caring concerns, such as the action of helping a significant other, and compare that to actions related to rights and justice concerns, such as the action of killing someone or stealing from someone. With respect to actions of the former type, while Americans generally tend to consider them to be under the

[6] It is worth noticing that the word 'obligation' seems to have some kind of part-whole polysemy, sometimes emphasizing what is obligatory, and sometimes emphasizing what is normative, as these terms are defined here. The word 'permissible' is used in the sense of discretionary as well, but this seems to be rather a case of generalized implicature (see Grice 1991).

[7] Since Wellman & Miller are referring to *moral* norms, this implies something more specific: what is *morally* right to do and what is *morally* wrong to do. For discussions of what may be at stake in this specificity, see Haidt *et al.*, 1993; Kelly & Stich, *forthcoming*; Nichols, 2004b; Shweder *et al.*, 1987; Turiel *et al.*, 1987.

scope of what is discretionary, Indians tend to consider them to be under the scope of what is normative. For example, given the scenario of an adult son who does not allow his elderly parents to live with him, even if he otherwise provides for their needs being taken care of, Americans would typically say that whether the son allows his parents to live with him is under the scope of what is discretionary ("It wasn't a life and death situation and their needs were being taken care of. Beyond that it's a personal choice."). In contrast, Indians would typically say that the son has a moral obligation to let his parents live with him (". . . It's a son's duty – birth duty – to take care of his parents. . . . the son has no business to ask his father to go away.").

Now, due to the fact that Americans have an individualistic conception of the self and Indians have a collectivist one, Americans and Indians have fundamentally different conceptions of what normative is. Because the individualist does not identify herself with collective concerns, social norms are typically seen as antagonistic to the realization of one's real preferences and their fulfilling is hence seen as unsatisfying – as far as possible, the ideal would be to live one's life under the scope of what is discretionary. Therefore, typically, Americans view social norms as a kind of coercive burden. Because the collectivist identifies herself with collective concerns, social norms are typically not seen as antagonistic to the realization of one's preferences and their fulfillment is hence seen as satisfying. Therefore, typically, Indians do not view social norms as a coercive burden. Take, for example, the scenario of a wife who, fulfilling her marital obligations, stays with her husband even after he has been severely injured and so cannot live up to the wife's marital expectations. Americans tend consider her fulfillment of duty as being in opposition to her individual desires and thus as unsatisfying ("She is acting out of obligation – not reasons like love. She has a sense of duty, but little satisfaction for her own happiness"). Indians, on the contrary, tend to assume that the woman is experiencing satisfaction in fulfilling her duty as a wife ("She will have the satisfaction of having fulfilled her duty. She helped her husband during difficulty.").

In his commentary, Charles Kalish proposes a specific way of interpreting the relation between psychological concepts and normative concepts, the two sides (or at least two important aspects of the two sides) of Wellman and Miller's notion of intentional responsive agency. He

accepts that many of the cognitive processes involved in normative evaluations may be independent of thinking about mental states, such as the decision of which norms should apply and the decision whether an action is consistent with or violates a norm. However, he thinks that fundamental for grasping the concept of norms is the understanding of their causal role in influencing intentional actions – their role as reasons. In other words, for Kalish, some psychological knowledge is an intrinsic part of the folk definition of norms:

> . . . obligations have their causal force as influences on intentional psychological processes. The claim here is that this causal role is central to the characterization of what an obligation is. If something is not understood to be a reason, then it is not an obligation. The concept of a reason is a prerequisite for concepts of obligations and other norms. Reasons are part of psychological knowledge. Thus a psychological understanding of the production of action is inseparable from normative knowledge. (Kalish, this volume, p. 201)
>
> . . . Norms (. . .) cannot be understood except as reasons for actions. Besides having a reason to do X, what does it mean that a person ought to do X? (Kalish, this volume, p. 201)

Free Will & Moral Responsibility

In his article "Folk Intuitions on Free Will," Shaun Nichols discusses folk intuitions concerning the free will problem *and* outlines the psychological mechanisms behind these intuitions.[8]

Let me start with folk intuitions. According to Nichols, the traditional problem of free will has two axes, both related to the concept of determinism – the idea that "every event is an inevitable outcome of the past conditions and the laws of nature." In the first axis, the query is whether human choices are so determined, that is, whether each token of a human choice could have been different, given its past conditions and the laws of nature. In the second axis, the query is whether the

[8] For important addenda to his article, see Nichols 2004a and Nichols & Joshua, *forthcoming*.

concept of moral responsibility is compatible with determinism, that is, whether one could be considered morally responsible if determinism were true. Therefore, in his article, Nichols discusses whether the folk concept of choice is deterministic or indeterministic *and* whether the folk concept of moral responsibility is compatible or incompatible with determinism.[9]

Let me start with the first axis. Endorsing what is, according to him, an implicit assumption in the current "Theory of Mind" literature, Nichols claims: "when engaged in the practical process of predicting and explaining behavior, people treat choice as deterministic."[10] To support this claim, he provides evidence about people's predictions of decisions of hypothetical perfect psychological duplicates – *different* persons who, immediately before making a decision, think, want, remember, see and feel exactly the same. People tend to say that if one of the duplicates decides to do something, the other duplicate will decide equally, which is consistent with a deterministic view of choice, since the exact same conditions existed before the decision.

---

[9] It is important to notice that the first axis of the problem is traditionally framed in compatibility terms too – whether determinism is compatible with free will in the sense of free choice (In this context, one should distinguish the problem of free will, which has two axes, and free will, which is primarily related to the first axis.). Of course, if there is an important folk sense of 'free will' that equals *indeterministic choice*, which is normally denied by compatibilists, free will in this sense is incompatible with determinism. Nichols avoids this other framing because "since there are many different notions of free will, the debate here often descends into squabbles about which kind of free will is under consideration." Nichols also recognizes that the notion of moral responsibility at stake in free will debates is not completely clear, which may lead to the same squabbles. He delimits the notion of moral responsibility of his interest as the one "tied to moral desert, blame, and retributive punishment."

[10] Thus, if the notion of indeterministic choice implies a certain kind of freedom (see previous footnote), current models of mind reading implicitly suppose that the "Theory of Mind" mechanism does not assign this kind of freedom to persons. The point here is not in competition with Wellman & Miller's aforementioned remark that the "Theory of Mind" literature tends to portray persons as autonomous agents, that is, in the context of discretionary choices, in the sense of what is discretionary that I characterized. This is because the type of autonomy implied by the notion of discretionary choice entails a different type of freedom – absence of coercive forces that go against one's real desires. Actually, this type of freedom may be considered to exist even in the context of normative influences at least by collectivists, in so far as they do not see norms as a coercive burden: "in such cultural communities many role-related duties are associated with individual satisfaction and experienced in *freely chosen* rather than controlling terms . . ." (Wellman & Miller, this volume, p. 43; the emphasis is mine).

However, Nichols claims that in other contexts people will tend to hold that choice is indeterministic. For example, children were presented with scenarios of physical events (e.g., a pot of water is put on a stove and boils) and of moral choices (e.g., Mary chooses to steal a candy bar). Then, they were asked whether, if everything in the world was the same right up until the physical event happens (the water boils) or until the moral choice is made (Mary chooses to steal a candy bar), the physical event *had to* happen or the moral choice *had to* be made. In answering these questions, children tend to say that the physical event had to happen but that the moral choice did not have to be made (see also Nichols 2004). Moreover, when adults were presented with the description of two universes – one in which global determinism is true and one in which only choice is indeterministic – and then were asked which of the two universes is most like ours, the great majority of subjects indicated the universe with indeterministic choice (see also Nichols & Knobe, *forthcoming*).

In sum, for Nichols, folk intuitions on whether choice is deterministic are mixed.

Now, turning to the other axis of the free will problem, Shaun reaches similar conclusions – sometimes people have incompatibilist intuitions, and sometimes people have compatibilist intuitions. His hypothesis is that, in the absence of emotional triggers, people will lean towards incompatibilist intuitions, but, when emotional triggers take place, they will lean towards compatibilist intuitions. For example, given the context of a deterministic universe, when adults are asked the abstract question of whether in this universe people can be fully morally responsible, they tend to respond as incompatibilists – people cannot be fully morally responsible – but when asked the concrete and emotionally laden question of whether in this universe a person is fully morally responsible for killing his family, they tend to respond as compatibilists – the person is fully morally responsible for killing his family (see also Nichols & Joshua, *forthcoming*).

Now, turning to the psychological mechanisms underpinning both kinds of mixed intuitions, Nichols raises the following hypotheses. Deterministic intuitions about choice come from the "Theory of Mind" mechanism, which seems plausible if one takes into account the fact that an indeterministic understanding of choice would not facilitate the explanatory and predictive function of this mechanism. Compatibilist intuitions about moral responsibility may be either the result of an affect-generated

performance error – in which case, compatibilist intuitions indicate an emotionally-driven misuse of the folk concept of moral responsibility – or the result of an affective competence – in which case, compatibilist intuitions indicate the proper application of the folk concept of moral responsibility, this concept being intrinsically linked to an affective competence.[11]

With respect to indeterministic and incompatibilist intuitions, Nichols approaches the problem in terms of their acquisition, that is, in terms of the acquisition of the concepts of indeterministic choice and incompatibilist moral responsibility. Contending that, for all we know, the traditional hypothesis in terms of introspection does not provide a good explanation for the acquisition of the concept of indeterministic choice,[12] and even less so for the acquisition of the incompatibililist concept of moral responsibility, he puts forward an alternative. He proposes that these concepts are acquired via the acquisition of the concept of obligation, of which we have solid evidence that children have a good grasp since very early in development. Suppose, *à la* Immanuel Kant, that the concept of obligation implies an indeterministic sense of could have done otherwise (*ought* implies *indeterministic can*): "The idea is that we can't be obligated to do the impossible, and if determinism is true, it is impossible for us ever to do other than we are determined to do. Thus, if we say that a person *ought* to have behaved differently, this implies that the person *could have done otherwise* (in an indeterministic sense)." Now, if children come to accept that the concept of obligation carries this implication, which may be true, they have a good reason to infer that choices are indeterministic. Furthermore, suppose that one is considered blameworthy for an action only if there is the normative expectation that one ought to have behaved differently. Now, if children also come to accept that this normative expectation is a necessary condition for the attribution of blame, which may be true, they have a good reason to infer that the concept of moral responsibility is incompatibilist – ought to have

[11] Nichols also raises a modularity hypothesis to explain these intuitions (see Nichols & Knobe, *forthcoming*, for a more complete discussion of this hypothesis).

[12] For a discussion of a broader range of hypotheses that could in principle explain the acquisition of the concept of indeterministic choice, see Nichols, 2004.

behaved differently implies indeterministic could have done otherwise and blameworthiness implies ought to have behaved differently, therefore blameworthiness implies indeterministic could have done otherwise and, reasonably, indeterministic choice.

In their commentaries, Paul Bloom, Eddy Nahmias and Charles Kalish call into question certain aspects of Nichols' interpretation of the folk concept of choice.

Paul Bloom challenges the (psychological duplicate) results showing deterministic intuitions about choice on methodological grounds and argues that humans tacitly believe in indeterministic choice, that is, they "hold an implicit view of human action that involves free will." Bloom suggests that, together with the idea that human consciousness is separate from the physical body and the idea that human selves continue to exist after biological death, the belief in indeterministic choice (in contrast with the belief in the deterministic nature of the physical world) is another consequence of human's intuitive dualism: "we are intuitive dualists, and we naturally explain the social-intentional domain in a very different way than the physical domain."

Eddy Nahmias suggests other interpretations of the results showing indeterministic intuitions about choice. On the one hand, he hypothesizes that what may be guiding most people's indeterministic intuitions is not something related to human choices in themselves (in opposition to physical processes), but rather something related to the complexity of the phenomena at hand:

> For simple processes, such as water boiling, holding fixed prior events may be considered sufficient to ensure the culminating event, but for complex processes, such as the weather, holding fixed prior events may *not* be considered sufficient to ensure later events. Some human decisions may be seen as complex in this sense and this might explain the pattern of responses Nichols got. (Nahmias, this volume, p. 219)

And he presents some experimental results where few people draw a distinction between human choices and physical processes with respect to indeterminism versus determinism (see also Turner & Nahmias, *forthcoming*), which is at least consistent with his hypothesis. On the other hand, he hypothesizes that part of what may be guiding people's denial that a universe with global determinism is similar to ours is the fact that

the scenario describing this universe may have been interpreted as a universe where fatalism, a stronger thesis than determinism, is true.[13]

Charles Kalish raises doubt about Nichols' developmental hypothesis about the acquisition of the concept of indeterministic choice. Nichols' hypothesis implies that the primary empirical input for the acquisition of the concept of indeterministic choice is the linguistic modal system; nonetheless, since the linguistic modal system is extremely ambiguous (normative terms have many non-normative usages – e.g., "Your skin ought not to be that shade of green." "That tree ought to have held your weight."), children cannot count on that system to identify what is an obligation: "Nichols might be right that an understanding of obligation leads to appreciation of volition, but a child cannot rely on the input to identify which things are voluntary, are reasons, or are obligations." (Kalish, this volume, p. 203)

In their commentaries, Manuel Vargas and Eddy Nahmias suggest that what may be at stake in people's incompatibilist intuitions on moral responsibility is the fear of a type of reductionism that would render our mental life epiphenomenal, instead of the fear of determinism in itself. Manuel Vargas indicates a study where subjects tended to attribute much less responsibility when a behavior is explained in physiological terms than when a behavior is explained in psychological or "experiential" terms. More to the point, Eddy Nahmias reports his study where subjects were given one of the following scenarios of alternate universes: a universe where the explanation of decisions and behaviors is couched in *neuro-chemical deterministic* terms, and a universe where the explanation is couched in *psychological deterministic* terms. Nahmias found that most subjects reading the former scenario denied that agents is this universe deserve to be given credit or blame for their actions, whereas most subjects reading the latter said that agents in this universe deserve to be given credit or blame for their actions.[14]

---

[13] He explicates the difference between determinism and fatalism as follows: "Determinism entails that $\Box[(Po \,\&\, L) \supset P]$ – i.e., necessarily, *given* the actual past state of affairs (Po) and the actual laws of nature (L), there is only one possible present state of affairs (P). But determinism does *not* entail (fatalism) that $\Box P$ (or that $\Box Po$ or $\Box L$) – i.e., that the actual state of affairs (or the actual past or laws) are necessary (could not be otherwise)."

[14] In his reply, Nichols addresses many of the criticisms pointed out here and even changes some of his original hypotheses.

Intentional Action & Moral Judgments

This part deals with the structure and function of the folk concept *intentional action* (or *acting intentionally*).[15] What is the structure of the folk concept of intentional action? Is its *primary* function the one of explaining and predicting behavior, which would make it a component of a "Theory of Mind" mechanism? Or is its *primary* function the one of establishing the moral status of an action and/or the blameworthiness or praiseworthiness of an agent, which would make it rather a component of moral reasoning?

Much of the discussion here is driven by some surprising results coming from Joshua Knobe's recent research (see, e.g., Knobe, 2003a, 2003b).

In their previous collaboration, Bertram Malle and Joshua Knobe, based on a range of experimental results, proposed a model of the folk concept of intentional action. In their model, an action A is considered intentional only if A is performed with skill and only if there is an intention to A – i.e., *skill* and *intention* are necessary conditions of the folk concept of intentional action (see Malle & Knobe, 1997, and Malle, this volume, for the complete model).[16] Thus, given a scenario of someone who tries to do something without the ability to do so and ends up doing so (e.g., tries to hit the bull's-eye and hits the bull's-eye without having any skill in using a gun), the model predicts that most people will say that the action (hit the bull's-eye) wasn't intentional, but a fluke. And, given a scenario where someone (e.g., the CEO of a company) *knows* that one of his intended actions (e.g., start a new program) will have some unintended side-effect (e.g., help the environment), and someone brings about the unintended side-effect (help the environment) via someone's intended action (start a new program), the model predicts that most people will say that the side-effect action (help the environment)

[15] It is important to keep in mind that, in this context, the words 'intentional,' 'intentionally,' and 'intentionality' are used to express the concept of a type of action. They refer neither to the property of "aboutness" of some mental states, nor to intention as a mental state.

[16] For various conceptual analyses of the concept of intentional action, see Adams, 1986; Bratman, 1987; Harman, 1976; Nadelhoffer, this volume; Mele & Moser, 1994; Mele & Sverdlik, 1997.

wasn't intentional, since it wasn't intended. Here are the scenarios of the two examples (see Knobe 2003a, 2003b):

> Jake desperately wants to win the rifle contest. He knows that he will only win the contest if he hits the bulls-eye. He raises the rifle, gets the bull's-eye in the sights, and presses the trigger.
> But Jake isn't very good at using his rifle. His hand slips on the barrel of the gun, and the shot goes wild . . .
> Nonetheless, the bullet lands directly on the bull's-eye. Jake wins the contest.

> The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, and it will also help the environment.'
> The chairman of the board answered, 'I don't care at all about helping the environment. I just want to make as much profit as I can. Let's start the new program.'
> They started the new program. Sure enough, the environment was helped.

The surprising results are that people's ascriptions of intentional action in these types of skill and side-effect scenarios do not always take into account skill and intention as necessary conditions and this seems to be related to the fact that people's ascriptions are in some way affected by moral judgments. Given the particular scenarios above, indeed most people say that the actions involved (hit the bull's-eye and help the environment) are not intentional – the predictions of their model are borne out. Nonetheless, if, in the first example, the action *hit the bull's-eye* is replaced with the action *kill someone*, and, in the second example, the side-effect action *harm the environment* substitutes for the side-effect action *help the environment*, most people in both cases now say that both actions are intentional – the predictions of the model do not hold.[17] Here are the new scenarios, where relevant actions are now considered intentional by subjects (see Knobe, 2003a, 2003b):

[17] These results have now been replicated several times, with various types of subjects, and various types of conditions (see references in this volume).

Jake desperately wants to have more money. He knows that he will inherit a lot of money when his aunt dies. One day, he sees his aunt walking by the window. He raises his rifle, gets her in the sights, and presses the trigger. But Jake isn't very good at using his rifle. His hand slips on the barrel of the gun, and the shot goes wild . . .
Nonetheless, the bullet hits her directly in the heart. She dies instantly.

The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.'
The chairman of the board answered, 'I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program.'
They started the new program. Sure enough, the environment was harmed.

It is worth pointing out two additional facts about current results. First, in these types of scenarios, subjects are also asked to judge the amount of praise or blame that the actors deserve for what they did, and the results show that subjects attribute *low* praise to the actors whose actions have a positive valence (hit the bull's-eye, help the environment), and *high* blame to the actors whose actions have a negative valence (kill someone, harm the environment). Thus, in these scenarios, there is an asymmetry in ascriptions of intentional action (ascription of *unintentional action* in the case of the positive action *versus* ascription of *intentional action* in the case of the negative action) that is analogous to the asymmetry in the ratings of praise and blame (*low* praise *versus high* blame).

Second, and this is related simply to the scenarios with side-effect actions, subjects are also asked whether the person *intended* to help or harm the environment. The results show that most subjects deny that the person intended to do so, in both conditions.[18] Therefore, while there is an asymmetry in ascriptions of intentional action between the help and the harm conditions, there is no asymmetry in terms of attributions of intention.

These results raise doubts not only on whether *skill* and *intention* are necessary conditions of the folk concept of intentional action but also

[18] This occurs with between-subject and within-subject designs.

on whether the primary function of this concept is the one of explaining and predicting behavior. Perhaps, the folk concept of intentional action is intrinsically linked to moral judgments. That's the question.

In his article "Intentionality, Morality, and Their Relationship in Human Judgment," Bertram Malle, focusing chiefly on results related to scenarios manipulating the skill variable, proposes to explain the aforementioned results without postulating any type of intrinsic relation between the folk concept of intentional action and moral judgments. He argues that two explanatory factors are relevant here. The first is the specific interplay between subjects' attention and the salience of the stimuli prompted by this type of scenario research:

> Making both intentionality judgments and blame judgments side by side but keeping them cognitively separate might either tax people's attentional resources or, relatedly, may not be the salient task participants discern in the vignettes. Attentionally engulfed by the evaluative information in the stories, people might conclude that their task is to take that evaluative information into account rather than make a conceptual or "technical" judgment of intentionality. This is particularly true when the evaluative information is so extreme (e.g., killing, sacrificing or saving many people's lives) that *not* responding to it would be seen as moral indifference. People adopt a more lenient criterion for intentionality judgments of evaluatively extreme actions not because of a conceptual dictate, but because of perceived demands to do so. (Malle, this volume, pp. 103-104)

He also emphasizes that this factor helps to explain not only the moral badness findings, but also one type of evidence that has been neglected by current explanatory models – the fact that Knobe also found that most people ascribe intentional action to cases where skill is absent but the action is an extremely positive one (e.g., a courageous soldier who luckily takes out a communication device and thereby saves many innocent lives – see Knobe, 2003a).

The second factor is people's sensitivity to what he calls "the scope of intentions," that is, "the range of actions and outcomes that would count as fulfilling the intention." The argument is that the vaguer the way an intention is specified, the more flexible are people's ascriptions of intentional action, due to the fact that a broader range of actions can fulfill a vaguer intention. Now, since in the skill scenarios used so far,

the content of the intention in the bad outcome condition is less specific than the content of the intention in the good outcome condition, this may facilitate, in the bad outcome condition, an ascription of intentional action that neglects skill as a necessary component.[19]

In their article "The Folk Concepts of Intention and Intentional Action: A Cross-Cultural Study," Joshua Knobe & Arudra Burra recognize that there have been fruitful attempts to explain the asymmetry in ascriptions of intentional action concerning bad and good side-effect actions. However, they argue that there has been no good explanation of the difference between ascriptions of intentional action and attributions of intention in this respect (i.e., of the fact that there is an asymmetry in ascriptions of intentional action but there is no asymmetry in attributions of intention). The main aim of their article is to consider the hypothesis that the folk concept of acting intentionally is radically unlike the folk concept of intention, because it is intrinsically linked to moral judgments:

> the concept of intention functions to facilitate predictions of behavior. But perhaps the concept of acting intentionally does not work like that; perhaps it should be understood primarily as a tool for making judgments about whether people deserve moral praise or blame for their behaviour. (Knobe & Burra, this volume, p. 114)

And they take the difference between ascriptions of intentional action (asymmetrical) and attributions of intention (symmetrical) as prima facie evidence that the two concepts are unrelated.

Moreover, they argue that 'intentionally' does not have the same meaning of the other manner adverbs in English that are derived from a noun. This exceptional character again suggests that the meaning of 'intentionally' is not derived from the meaning of 'intention.' According to Knobe & Burra, the relation between 'intentionally' and 'intention' is rather like the relation between 'rationally' and 'ration' – "just two separate words that happen to be morphologically related."

[19] In his article, Malle also provides some experimental evidence confirming the relevance of this factor.

Finally, and more fundamental to their argument, they show some parallel evidence coming from a study with Hindi language. They gave the same side-effect scenarios to Hindi-speaking students, and used the Hindi word 'jaan-bujhkar,' which apparently is the word for *intentionally*, to frame the intentionality question. They found the same asymmetry in ascriptions of intentional action – most people said "that the chairman harmed the environment *jaan-bujhkar* but not that he helped the environment *jaan-bujhkar*." They also asked a question using the related word 'jaan,' which, in Hindi, express *knowledge*, instead of *intention*. They did not find any difference in attributions of knowledge between the help and harm conditions. This absence of asymmetry "indicates that the word 'jaan-bujhkar' is not simply an adverbial form of 'jaan.'" So, Knobe & Burra hypothesize that:

> (. . .) the meaning of the word 'intentionally' is a kind of primitive. We do not understand the meaning of 'intentionally' by understanding the meanings of its component morphemes and then understanding how they fit together to form the meaning of the whole. Rather, we have an independent concept of acting intentionally (distinct from our concept of intention), and we understand the meaning of 'intentionally' by understanding that it expresses this concept. The mind includes certain mechanisms for determining whether or not a given behavior was performed intentionally, and it seems likely that these mechanisms make use of various other concepts. But there is a big difference between (a) the relatively banal claim that we use various other concepts to determine whether or not a behavior was performed intentionally and (b) the more controversial claim that the word 'intentionally' can actually be *defined* in terms of other concepts. So, for example, suppose that we have an innate 'moral faculty' (. . .) and that this faculty can determine whether or not behaviors were performed intentionally. When we learn the meaning of the word 'intentionally,' we might simply be learning to map that word onto a concept that is already being used by the moral faculty. On this model, the moral faculty might be using various other concepts to determine whether or not a behavior was performed intentionally, but the language faculty does not contain a definition of 'intentionally' in terms of other concepts. (Knobe & Burra, this volume, pp. 123-124)

In his article "Desire, Foresight, Intentions, and Intentional Actions: Probing Folk Intuitions," Thomas Nadelhoffer explicates various philosophical models of the relations between the concepts of desire, fore-

sight, intention and intentional action. In this context, he interprets these models as "predictions about how laypersons would respond to particular cases." Then, he presents empirical results dealing with laypersons' intuitions on classic scenarios of side effects discussed in the action theory literature. The findings, which in general are consistent with previous results, are as follows. Ascriptions of intentional action are sensitive to the degree of foresight an agent has in relation to the possible side effects of her actions – the higher the perceived likelihood of the side effect, the more likely people are to judge that the side effect was brought about intentionally. Second, in cases of side-effect scenarios, people tend to ascribe intentional actions more than they tend to attribute intentions. Finally, people often judge that an agent intentionally brought about a given side effect even though she did not want to do so, and this is especially the case when the agent is deemed to blame for bringing about the side effect.

In his commentary, Fred Adams reiterates his pragmatic account of these results and extends it to the cross-cultural data provided by Knobe & Burra. On his pragmatic account, subjects answering these scenarios are using the pragmatic meaning of the word 'intentionally,' instead of its literal meaning. They are asserting or denying intentionality not in terms of the descriptive content of the literal meaning of 'intentionally,' but as a way of implicating blame or avoiding an implication of praise. So, the asymmetry in judgments of bad and good side effects is due to the fact that they want to blame the CEO who does not care about harming the environment, and they do not want to praise the CEO who does not care about helping the environment. According to Adam, the same rationale can be easily applied to the cross-cultural results:

> Hindi speakers want to discourage one for harming the environment "intentionally" (or knowingly, given the root derivation of '*jaan-bujhkar*' from '*jaan*'). They know the best way to do this is to use the pragmatic weight of the term '*jaan-bujhkar*.' Similarly, they don't want to praise him for helping the environment knowingly, since he says (in Hindi) the equivalent of "I don't care at all about helping the environment. (Adams, this volume, p. 265)

In his contribution, Charles Kalish claims that the side-effect results do not demonstrate that "the same mental process may be described as acting intentionally or not depending on our evaluation of the outcome."

Rather, subjects are taking into account the actor's *beliefs* about the positive and negative value of their actions – in negative outcome scenario, the CEO considers and rejects a reason not to implement the program (to harm the environment); in the positive outcome, the executive recognizes an additional reason to go ahead with the program (to help the environment) (cf. Harman's comments). Furthermore, Kalish claims that pragmatic communicative principles such as relevance may be guiding subjects' loose ascriptions of intentional action – "communicative goals might determine whether an action was "intentional enough" to count as acting intentionally."

In their contribution to the debate, Liane Young, Fiery Cushman, Ralph Adolphs, Daniel Tranel and Marc Hauser report results showing that the asymmetry in Knobe's side-effect scenarios arises even with subjects who have sustained damage to the ventromedial prefrontal cortex (VMPC) and who, for this reason, display severely compromised emotion processing. Based on this finding, Liane *et al.* suggest that normal emotional processing is not responsible for the observed asymmetry in ascriptions of intentional action and hence does not mediate the relationship between an action's moral status and its intentional status. In other words, their results go against the hypothesis that, in the case of the harm condition, people are ascribing intentionality because of an immediate emotional reaction to the fact that the CEO allows the environment to be harmed.

In his commentary, Gilbert Harman raises various doubts about Knobe & Burra's hypotheses. According to Harman, people's ascriptions of intentional action seem to be "sensitive to whether there is a (prima facie or default) *reason* against doing what is done as a foreseen side effect." (see also Harman, 1976) Furthermore, this reason needs not be a *moral* reason, and, if folk psychology includes this general notion of reason, as he believes, it does include a notion of acting intentionally. Harman also questions Knobe & Burra's analyses of the relation between manner adverbs and nouns in English, and their construal of *intentionally* as a primitive concept.

In his commentary, Alfred Mele accomplishes a comprehensive appraisal of current empirical results concerning the folk concept of intentional action. Assuming that so far we do not have sufficient data to be confident in any specific analysis of the folk concept of intentional action,

Mele envisages what would be the shape of an analysis of such a concept, if one takes current results at face value. According to him, a face-value theorist should start making a general division between actions that are side-effect actions and actions that are not side-effect actions, then analyze the folk concept of intentional action of each of these types separately, and finally build a disjunctive analysis of the folk concept of intentional action out of these analyses. In his commentary, Mele, as a face-value theorist, arrives at the contours of such a disjunctive analysis and proposes various research paths to improve the quality of the data.[20]

REFERENCES

Adams, F.
1986 Intention and intentional action: The Simple View. *Mind and Language, 1*, 281-301.

Alicke, M.
2000 Culpable control and the psychology of blame. *Psychological Bulletin* 126/4: 556-574.

Bloom, P.
2004 *Descartes' Baby*. New York: Basic Books.

Bratman, M.
1987 *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

Churchland, P.
1979 *Scientific Realism and the Plasticity of Mind*. Cambridge: Cambridge University Press, 1979.

Darley, J. M. & Shultz, T. R.
1990 Moral Rules: their content and acquisition. *Annu. Rev. Psychol.* 41:525–56.

Grice, P.
1991 Logic and Conversation. In S. Davis (ed.) *Pragmatics – A reader*. Oxford: Oxford University Press.

Haidt, J., Koller H., & Dias M.
1993 Affect, Culture and Morality, or Is it wrong to eat your dog? *Journal of Personality and Social Psychology*, Vol. 65, No. 4, 613-628.

Harman, G.
1976 "Practical Reasoning." *Review of Metaphysics* 79: 431-463.

Harris, P.
2000 *The Work of the Imagination*. Oxford: Blackwell Publishers.

[20] In their reply, Malle (with Gluglielmo) and Knobe address many ideas raised by the commentaries.

Hauser, M.
Forthcoming *Moral minds: The unconscious voice of right and wrong*. NY: Harper Collins.
Jackendoff, R.
1999 The natural logic of rights and obligations. In R. Jackendoff, P. Bloom, and K. Wynn (eds.), *Language, Logic, and Concepts – essays in memory of John Macnamara*, Cambridge: The MIT Press.
Kalish, C. W.
2005 Becoming status conscious: Children's appreciation of social reality. *Philosophical Explorations*, 8, 245-263.
Kalish, C. W., & Shiverick, S. M.
2004 Children's reasoning about norms and traits as motives for behavior. *Cognitive Development*, 19, 401-416.
Kalish, C. W., Weissman, M., & Bernstein, D.
2000 Taking decisions seriously: Young children's understanding of conventional truth. *Child Development*, 71, 1289-1308.
Kelly, D. & Stich, S.
Forthcoming Two theories about the cognitive architecture underlying morality. In P. Carruthers, S. Lawrence & S. Stich (eds.) *The innate mind – structure and content*. Cambridge: Cambridge University Press.
Knobe, J.
2003a Intentional Action and Side-Effects in Ordinary Language. *Analysis* 63: 190–93.
2003b Intentional Action in Folk Psychology: An Experimental Investigation. *Philosophical Psychology* 16: 309-324.
Leslie, A.M.
1994 ToMM, Toby, and Agency: Core Architecture and domain specificity. In *Mapping the Mind*, L. Hirschfeld & S. Gelman (eds.). Cambridge: Cambridge University Press.
Malle, B. F., Moses, L. J., & Baldwin, D. A. (Eds.).
2001 *Intentions and intentionality: Foundations of social cognition*. 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*, 563-580.
Malle, B. F., & Knobe, J.
1997 The folk concept of intentionality. *Journal of Experimental Social Psychology, 33*, 101-121.
McCauley, R.
1986 Intertheoretical relations and the future of psychology. *Philosophy of Science*, 53.
Mele, A. and P. Moser
1994 Intentional Action. Nous 28: 39-68.
Mele, A. R., & Sverdlik, S.
1996 Intention, Intentional Action, and Moral Responsibility. *Philosophical Studies*, 82, 265-87.
Morton, A.
2004 *The importance of being understood: folk psychology as ethics*. Routledge.

Nadelhoffer, T.
2005 Skill, luck, and intentional action. *Philosophical Psychology*, 18:3, 343-354.
Nahmias, E., Morris, S., Nadelhoffer, T. & Turner, J.
2005 Surveying Freedom: Folk Intuitions about Free Will and Moral Responsibility. *Philosophical Psychology*, 18: 561-584.
Nichols, S.
2004a The Folk Psychology of Free Will: Fits and Starts. *Mind & Language*, 19, 473-502.
2004b *Sentimental Rules – On the natural foundations of Moral Judgement*. New York: Oxford University Press.
Nichols, S. & Knobe, J.
Forthcoming Moral Responsibility and Determinism: Empirical Investigations of Folk Concepts.
Nichols, S. & Stich, S.
2003 *Mindreading*. Oxford: Oxford University Press.
Nunez, M., & Harris, P. L.
1998 Psychological and deontic concepts: Separate domains or intimate connection. *Mind & Language, 13*, 153-170.
Shaver, K.G.
1985 *The attribution of blame: Causality, responsibility and blameworthiness*. New York: Springer.
Shweder, R. A., Mahapatra, M. & Miller, J.
1987 Culture and moral development. In J. Kagan and S. Lamb (eds.) *The emergence of morality in young children*. Chicago: University of Chicago Press.
Turiel, E.
1983 *The development of social knowledge: Morality and convention*. New York: Cambridge University Press.
1998 *The development of morality*. In W. Damon & N. Eisenberg (Eds.), Handbook of child psychology, 5th ed.: Vol 3, social, emotional, and personality
Turiel, E., M. Killen, and C. Helwig
1987 Morality: Its structure, functions, and vagaries. In J. Kagan and S. Lamb (eds.) *The emergence of morality in young children*. Chicago: University of Chicago Press.
Turner, J. and Nahmias, E.
Forthcoming Are the folk agent-causationists? *Mind and Language*.
Weiner, B.
1995 *Judgments of responsibility*. Guilford.
Wellman, H.
1990 *The child's theory of mind*. Cambridge: MIT Press.
Wellman, H. M. & Miller, J. G.
Forthcoming Including deontic reasoning as fundamental to theory of mind.
Woolfolk, Robert L., John Doris, and John Darley
Forthcoming Identification, Situational Constraint, and Social Cognition: Studies in the Attribution of Moral Responsibility. *Cognition*.