Mind and Morality

Peter H. Kahn, Jr.

University of Washington


I thank Stephanie Carlson and Charles Helwig for comments on an earlier version of this paper.

Author's address: Department of Psychology, Box 351525, University of Washington, Seattle, WA 98195-1525. Email: pkahn@u.washington.edu.
It is puzzling that the fields of Theory of Mind and Moral Development have remained largely divided over the years. Is the reason mostly due to their different starting points, or to an academic culture that rewards "novelty?" Or, as Sokol, Chandler, and Jones (chap. 2) suggest, are there fundamental theoretical disagreements that make bridging these two fields difficult? In this commentary, I provide some theoretical, historical, and personal perspective on these questions. Too, by discussing the chapters I hope to further the integrative goals of this timely volume.

**ToM and Computational Conceptions of the Mind**

Sokol et al. argue that ToM has "promoted an impoverished conception of agency, that, in the end, proves to be incompatible with our folk conceptions of morality." I would like to unpack this idea, and offer a slightly different conclusion.

First, let us get a sense of why people like ourselves ("folk") might have conceptions of morality. Imagine three situations. In the first situation, a man is walking in the woods, alongside some craggy cliffs. A boulder dislodges from the ledges above, plummets down, and kills the man. As unfortunate as this situation is for the man, we would not normally say that the boulder acted immorally. In the second situation, a man is walking in the woods, and a mountain lion pounces out from the trees and kills the man (and feeds her baby lion cubs on him). Again, from the man's standpoint this situation is most unfortunate. But we would not normally say that the lion acted immorally, even though the lion is a biological animal (and some readers might even say has intentions and desires). In the third situation, a man is walking in the woods, and a bandit, with a knife in hand, steals the man's money, and on parting kills him. It is only in this latter situation that we would normally say something unethical or immoral has occurred. Why do we say this? Presumably it is because we believe that morality involves not only intentions and desires, but intentions and desires of a certain type: for example, desires for human well
being, and for fairness and justice in social relationships. We also believe that morality involves a significant measure of free choice, such that people can be held responsible for their actions.

Sokol et al. point out that the psychological fields have a long history of dismissing such folk conceptions of morality. B. F. Skinner, for example, who for decades helped chart the behavioristic course of psychology, believed that moral (or immoral) acts have no more "moral" (or "immoral") epistemic standing than any other human act. As Skinner (1971) wrote: "Relevant social contingencies are implied by 'You ought not to steal,' which could be translated, 'If you tend to avoid punishment, avoid stealing,' or 'Stealing is wrong, and wrong behavior is punished.' Such a statement is no more normative than 'If coffee keeps you awake when you want to go to sleep, don't drink it'" (p. 114). It is for such reasons that Skinner titled the above referenced book *Beyond Freedom and Dignity*. He could have as equally well titled this book (or another) *Beyond Morality*.

In recent decades, scientific psychological theories have similarly drawn on nonmoral conceptions of the human mind, particularly by employing mechanistic metaphors. Cosmides, Tooby, and Barkow (1992), for example, write of "information-processing mechanisms situated in human minds" (p. 3), and of the brain as "a computer made out of organic compounds" (p. 8). Pinker (1997) argues "that the mind is a system of organs of computation designed by natural selection to solve the problems faced by our evolutionary ancestors in their foraging way of life" (p. x). Dawkins (1976) says: "We are survival machines – robot vehicles blindly programmed to preserve the selfish molecules known as genes" (p. ix).

Such analogies between the human mind and machine are particularly perplexing when we broach the subject of morality. By most accounts machines are not moral agents. If we are machines, wherein lies our moral agency? Of course, at times it may be useful to think of the human mind as if it were a piece of computational machinery. But, as Searle (1990) notes, while it can be useful to model water molecules with ping-pong balls in a bathtub, no one jumps into a
bath tub of ping-pong balls expecting to get wet. Similarly, in our computational modeling we should not confuse mind with machine. Sokol et al.’s point is that ToM does just that.

Are Sokol et al. correct? I think the answer depends on what one reads in the ToM literature, and – even then – whether one takes the words at face value. For example, Gopnik, Meltzoff, and Kuhl (1999) write:

...babies are a kind of very special computer. They are computers made of neurons, instead of silicon chips, and programmed by evolution, instead of by guys with pocket protectors. They take input from the world, the flickering chaos of sensations, and they (and therefore we) somehow turn it into jokes, apologies, tables, and spoons. Our job as developmental psychologists is to discover what program babies run and, someday, how that program is coded in their brains and how it evolved. (p. 6)

Later, Gopnik et al. (1999) write: "The Big Idea, the conceptual breakthrough of the last thirty years of psychology, is that the brain is a kind of computer. That's the basis of the new field of cognitive science" (p. 21). Yet even when ToM researchers employ strong computational language, it is hard to believe that they would vouch for its reductionist and mechanistic implications. After all, their research publications are filled with rich, nuanced accounts of the child's interpretative and intentional stance. Such accounts seem incompatible with a conception of the child as a programmed, nonmoral automaton.

That said, I think we need to heed Sokol et al.’s message. The ToM metaphor of the mind as computer needs to be more carefully articulated to better support concepts of meaning, morality, and free will. The ToM computational framework also needs to account for how individuals can shape, from an ethical stance, cultural practices. These are not easy tasks. Thus one of the strengths of this current volume is that it draws on the moral-developmental literature, and especially the "domain-specific" approach to moral development, to help ground the moral underpinnings of ToM.
The Domain-Specific Approach to Moral Development: Historical and Personal Reflections

The domain-specific approach to moral development emerged more than 2 decades ago, in part as a correction to Kohlberg's theory which views moral development in terms of increasing differentiation. According to Kohlberg (1969, 1971), through development moral judgments pull away from personal concerns (stages 1 and 2) and then conventional concerns (stages 3 and 4) before achieving the generalizable and universal features of principled moral reasoning (stages 5 and 6). Yet, according to domain theorists, one problem with such an account is that it underestimates the ways in which even young children construct distinctly moral judgments.

For example, suppose we ask a 7-year-old child "Is it all right or not all right for Sally to push Joan off the swing because Sally wants to swing?" The child might initially provide a Kohlbergian-looking stage 1 response: "No, that's not all right because Sally could get in trouble." But then say we counter probed with any number of different questions. We ask the child: "Let's say Sally wouldn't get in trouble, would it then be all right or not all right?" Or we ask: "Let's say the rule at Sally's school says that if you want to push, it's up to you; would it then be all right or not all right?" Or we ask: "Let's say that in another school [or country], the way they did things there is it's up to the child whether he or she wants to push; would it be all right or not all right for children in that school [or country] to push in this sort of situation?" What domain theorists have found is that for all such questions children typically say, "it's still not all right"; and children typically base their judgments on considerations of neither fear of punishment nor conventional practices, but human welfare (e.g., "because Joan could get hurt") or fairness (e.g., "it's not fair, because Joan was swinging first, and Sally should just wait her turn"). More formally, Turiel (1983) says that such questions (that elicit what he calls criterion judgments) establish conceptual boundaries for the moral domain, based on the criteria of prescriptivity, rule contingency, and generalizability (and supporting welfare and fairness justifications). Based on these criteria,
domain theorists have provided convincing evidence that even young children bring forward
distinctively moral judgments (see Smetana, 1995; Tisak, 1995; Turiel, 1998; for reviews).

During the early 1980's, I was beginning my graduate work in moral development. Part
of what I found particularly powerful about the domain-specific approach was that it sought to
place moral judgments within a larger developmental framework. For example, in his 1983
volume, titled *The Development of Social Knowledge*, Turiel lays out a research agenda:

The analyses and documentation presented in this volume provide the foundation for
extending...the domain-specific interpretation of structure and development to a more
general theoretical framework regarding social development...On the basis of the
assumption that the individual's social world includes other persons, relations between
persons, and institutionalized systems of social interaction, it is proposed that the child's
structuring of the social world revolves around three general categories. These are (1)
concepts of persons or psychological systems (the psychological domain), (2)
conceptions of systems of social relations and organizations (the societal domain) of
which convention is but one component, and (3) prescriptive judgments of justice, rights,
and welfare (the moral domain). (pp. 3-4)

Thus, according to Turiel, moral judgments need to be understood within the context of 3
domains of social knowledge: the moral, social-conventional, and psychological.

At the time, the psychological domain (which involved "concepts of persons or
psychological systems") was understood to include children's concepts of agency and
intentionality: a version of what has become known as a Theory of Mind. Thus I thought domain
theorists were particularly well poised for the integration of psychological, social, and moral
reasoning, for creating a unified theory, if you will, of social development. In subsequent years,
however, the psychological domain became blurred with the "personal domain": concepts that
pertain to individual authority, and which focus on actions (e.g., one's choice of friends and
recreational activities) that often lie outside the realm of justifiable social regulation (Nucci,
1981, 1996). I was puzzled at the time, wondering if perhaps the personal domain could be thought of as a subset of the psychological domain; but I didn't worry unduly because I figured it would all be straightened out in due course. But it hasn't been. While Nucci (2001) and others continue to make important contributions to the field based on analyses of personal reasoning, an account of the psychological domain has not been developed.

During this entire period, starting in the 1980's, I and others from the domain perspective were certainly aware that another area in psychological research was starting to emerge, called Theory of Mind. To be honest, at the time, I (and others) thought ToM was emerging from too narrow an empirical basis (e.g., the false-belief task) and driven by what seemed quick publishable experiments. And, like Sokol et al., I thought ToM seemed rooted to a computational conception of the mind that would prevent serious engagement with and theorizing about the moral life. Since then, ToM has perhaps mined most of what is possible from the false-belief paradigm, and is now broadening into areas of biology, culture, and, as evidenced from this volume, moral development. What can the shape of such research look like?

New Directions for Theory of Mind and Moral Development

The contributors to this volume take up this question head on. Baird (chap. 3) stakes out a key area of research for the field today: investigating children's ability to consider intentions in judging the moral quality of others' actions. This distinction between moral intention and action is crucial in moral theorizing, our courts of law, and of course our common judgments. After all, say I told you: "Bob spends every Saturday at his aunt's house, helping her with chores." You might think "oh, what a wonderful person Bob is." Then I tell you, "Bob only helps his aunt because the aunt is rich and old and close to death, and Bob hopes to inherit boodles." Then you might think, "Well, Bob's not quite the moral guy I thought he was." In other words, it matters why we do the things we do.
Such investigations into conceptions of moral intentions and consequences have a long history in the moral-developmental literature (Piaget, 1932/1969). Against this backdrop, Baird sought to vary intentions ("good" vs. "bad") while keeping the act (e.g., turning on a garden hose) identical. Her results suggest that both 4- and 5-year-olds can evaluate and punish differently on the basis of the actor's intentions. To be brief, I suspect Baird's conclusion is correct (cf. Helwig, Hildebrandt, & Turiel, 1995), but I am less convinced by the methodical design of the study. Baird sought to use a "neutral" stimulus (e.g., turning on a garden hose) that could then be paired with a "good" intention (e.g., watering the garden) and "bad" intention (e.g., ruining a sibling's sandcastle). But it is not clear to me that the "neutral" action can be severed from the "good" and "bad" actions that follow. In other words, when a child is reasoning about the events, it seems likely he or she is considering not only two different intentions and the "neutral" action but two different consequences, and that those consequences are wrapped up in the evaluation of the seemingly neutral action. Granted, Baird did not provide children with information about consequences; thus for my interpretation to be correct, children would have to be inferring or creating consequences on their own. But if the children did, then Baird's experimental design may not allow for the precision that she seeks in her interpretation.

In future investigations, I would also add two other pointers from the moral-developmental side. If one wants to examine moral reasoning, it is important to employ moral stimuli. Is helping one's mother water a garden moral? Perhaps it is in a morally "discretionary" sense (Kahn, 1992, 1997, 2002) but not in the classic moral sense by which domain-theorists understand the term. Similarly, when seeking moral evaluations, it is usually not enough to garner judgments of "good" vs. "bad." After all, people also make judgments that many types of non-moral violations are bad (or good). For example, many people would say that it is "bad" to eat a baked potato with your fingers at a fancy restaurant, but only because of conventional (not moral) expectations. It is for this reason that domain theorists employ criterion judgments.
Like Baird, Moore and Macgillivray (chap. 4) stake out a rich area for investigation. They are interested in how a child's psychological understanding of self and others (ToM) can help regulate the child's own moral behavior. Moore and Macgillivray report on a series of studies wherein they employed "delay of gratification" tasks. For example, one task involves a game, wherein a child was offered one sticker now or two stickers at the end of the game. A second (no-cost sharing) task involved a choice between one sticker now for self and two stickers now shared between the two players. In other manipulations, Moore and Macgillivray investigated what they call "simple sharing with cost" (two stickers now or one a piece now), "delayed sharing" (two stickers for self now or one each later), and "future-oriented prudence" (one sticker for self now or two for self later). Through such manipulations, Moore and Macgillivray sought to understand the sources of variation for sharing (e.g., developmental or individual differences), and the relationship between prosocial behavior and prudence. Taken together, they also discuss how prudence and prosocial behavior relate to ToM.

Moore and Macgillivray ask good questions, and are sensitive to the complexity of their results. I would just ask one further question. How does all of this (prudence, prosocial behavior, and ToM) relate to moral development? A likely answer is that "prosocial behavior" is the moral component. But some time ago it was clear to many in the field that prosocial behavior is a broad term that can refer to both moral and non-moral acts (Smetana, Bridgeman, & Turiel, 1983). Moreover, we need to remember Baird's earlier point: to establish the moral basis of an act, the moral description needs to include both behavior and reasoning. Consider another example: Imagine if Jane walks past a homeless person on the street who asks her for money. Is that immoral behavior? Nonmoral behavior? Well, what were Jane's reasons? Maybe she saw the homeless man drinking alcohol and worried that giving the man money would lead to his further deterioration. Maybe she actually knew the man, and believed that he was making a choice to be
on the street, and could successfully gain employment if he so chose. Or say Jane does give the homeless man money. Maybe she does so only out of fear that otherwise the homeless man will assault her. Or maybe she believes that a person shouldn't walk past another human being in need without somehow recognizing their dignity; and in this context money is the mechanism for conveying her recognition. The point is that children, too, bring complex assessments of context and moral judgments to bear in their social and moral reasoning – even in laboratory contexts. Thus, it would be good (in the nonmoral sense, of course) to assess such reasons, and then to integrate them into further analyses of social competence and ToM.

Drawing on the domain-specific approach, Wainryb (chap. 1) provides empirical support for the proposition that children can arrive at different moral decisions because of a dispute not over their moral beliefs but factual beliefs. An important implication follows for cross-cultural investigations. To explicate how, consider Shweder, Mahapatra, and Miller's (1987) data that shows that devout Hindus believe that it is immoral for menstruating women to sleep in the same beds as their husbands. When Turiel, Killen, and Helwig (1987) reexamined the data, they found that devout Hindus believed that menstrual blood is poisonous and harmful to husbands. In such a situation, devout Hindus and traditional Westerners presumably share the moral belief that wives should care for the well being of their husbands. Where they differ is in the factual belief about what causes harm. Thus Wainryb's research helps check anthropologists (and many lay people, as well) who too often jump to the conclusion that when cultures differ in their practices that the cultures also differ substantively in their moral judgments.

Wainryb also offers some provocative connections to ToM, since, she, too, is interested in "false beliefs." Indeed, Wainryb suggests that false beliefs may be better categorized in at least two ways. In one way, a false belief can be understood as a mistaken belief. For example, her 4-year old daughter, Julia, held a mistaken belief that a person can see every type of bug ("How can there be bugs that no one can see?"). In a second way, the false belief can be understood as a belief that others can have legitimate grounds for an alternative interpretation of reality. For
example, while a person may not believe that spanking a child is pedagogically desirable, he or she can recognize that others might, and (with limits) perhaps give some ground to such behavior. Like Sokal et al. and their colleagues (Carpendale & Chandler, 1996; Chandler & Lalonde, 1996), Wainryb has found that children's understanding that others can have alternative interpretations comes later than the understanding that others can have mistaken beliefs.

In such ways, Wainryb offers a powerful and innovative approach to investigating social judgments. Indeed, I draw on her work frequently in my own research and theorizing (Kahn, 1999; Kahn & Lourenço, 1999). Yet over the years I have also wondered whether one can completely disentangle moral beliefs from factual beliefs. As a case in point, consider the factual belief of a 19th century American, R. R. Cobb, when he was addressing the issue of slavery: "A state of bondage, so far from doing violence to the law of nature, develops and perfects it; and that, in that state (the Negro) enjoys the greatest amount of happiness, and arrives at the greatest degree of perfection, of which his nature is capable" (quoted in Spiegel, 1988, p. 39). Let us imagine we challenged Cobb. We might say: "Mr. Cobb, you are an immoral personal." Cobb might respond: "You are mistaken. I care very much for the welfare of human beings, and always treat humans with respect and dignity. But the fact is that Negroes are not fully human, and thus (like animals) are not accorded the same moral standing as humans." With such reasoning at hand, and Wainryb's distinction employed, we could then be left saying: "Mr. Cobb, isn't it interesting that we actually share the same moral beliefs, and only differ regarding our factual assumptions." While there is something right about this statement, there seems to me something wrong about it, as well, insofar as some factual beliefs appear so morally laden that to get them wrong implicates one's moral life.

Wainryb is aware of such potential difficulties, and moves deftly to address them. For example, she recognizes that people may on occasion deliberately or unconsciously conceal what they believe to be true, as a means to advance personal goals or to evade inner conflict. Yet
Wainryb, I think correctly, argues that such rationalization does not characterize the central unit of analysis in the moral life.

Wainryb also responds to the potential charge that her position could be construed as morally relativistic. She writes, for example: "We do indeed argue that because subjective construals of reality can transform meanings of 'facts' any act can be judged as immoral, but only insofar as its 'facts' are understood to signal the presence of intentional harm. This is not akin to saying, as moral relativists do, that the relation between act and moral judgment is arbitrary" (cf. Wainryb, 2000). Yet here I am not convinced that Wainryb can give up so much initial ground (that "any act can be judged as immoral") and still maintain (by means of an individual's perception or judgment of "intentional harm") a firm non-arbitrary connection between act and moral judgment. After all, individuals sometimes differ in their perception or judgment of when intentional harm occurs. Moreover, what checks the "Mr. Cobb" scenario above – where construals of reality (factual beliefs) transform seemingly immoral acts (slavery) into non-moral acts? In such ways, there is a greater amount of moral relativism embedded in Wainryb's current position than most domain theorists (Wainryb included) have proposed in the past.

Conclusion

From this small volume lies the potential for the fields of ToM and Moral Development to enrich one another. ToM can help moral-developmental domain researchers fulfill the original promise of the psychological domain, uncovering children's concepts of persons and psychological systems. From such work, it would then be possible to develop a more unified theory of social development. Conversely, domain theory can help move ToM toward non-reductionistic theorizing and research. But for this move to succeed, moral beliefs need to be viewed within ToM not as epiphenomenal (as mere "folk" conceptions, in the way that "natives" might have folk conceptions that the world is flat), but as playing an epistemically valid and psychologically authentic role in human lives. Toward this end, it always good to pay attention to
the language of one's characterizations. The title of this volume, for example, is *The Interface between Children's Theories of Mind and Socio-Moral Development*. I suggest that our thoughtful contributors have established important dialog (not an "interface") between the two fields. Perhaps fruitful empirical collaborations will follow.
References


Headnote: The author suggests that Theory of Mind can help the Moral Developmental field fulfill the original promise of the psychological domain, uncovering children's concepts of persons and psychological systems. Conversely, Moral Developmental theory can help ToM move toward non-mechanistic moral theorizing and research, and toward an account of how individuals can shape, from an ethical stance, cultural practices.

Description of Chapter (for Table of Contents): An examination of why the fields of Theory of Mind and Moral Development have remained largely divided over the years, and how they can now enrich one another.

Author's Affiliation: Peter H. Kahn, Jr. is Research Associate Professor of Psychology at the University of Washington. He is also Adjunct Research Associate Professor in the Information School at the University of Washington and Co-Director of the Mina Institute in Covelo, California.