Piagetian Roboethics via Category Theory: Moving Beyond Mere Formal Operations to Engineer Robots Whose Decisions are Guaranteed to be Ethically Correct

Selmer Bringsjord, Joshua Taylor
Trevor Housten, Bram van Heuveln
Konstantine Arkoudas, Micah Clark
Rensselaer AI & Reasoning (RAIR) Lab
Department of Cognitive Science
Department of Computer Science
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA

Ralph Wojtowicz
Metron Inc.
1818 Library Street
Suite 600
Reston VA 20190 USA

I. Introduction

This is an extended abstract, not a polished paper; an *approach* to, rather than the results of, sustained research and development in the area of roboethics is described herein. Encapsulated, the approach is to engineer ethically correct robots by giving them the capacity to reason *over*, rather than merely *in*, logical systems (where logical systems are used to formalize such things as ethical codes of conduct for warfighting robots). This is to be accomplished by taking seriously Piaget's position that sophisticated human thinking exceeds even abstract processes carried out *in* a logical system, and by exploiting category theory to render in rigorous form, suitable for mechanization, structure-preserving mappings that Bringsjord, an avowed Piagetian, sees to be central in rigorous and rational human ethical decision-making.

We assume our readers to be at least somewhat familiar with elementary classical logic and category theory. Introductory coverage of the former subject can be found in [1], [2];[1] such coverage of the latter, offered from a suitably computational perspective, is provided in [3]. Additional references are of course provided in the course of this document.

II. Piaget's View of Thinking

Many people, including many outside psychology and cognitive science, know that Piaget seminally — and by Bringsjord's lights, correctly — articulated and defended the view that mature human reasoning and decision-making consists in processes operating for the most part on formulas in the language of classical extensional logic (e.g., see [4]).[2] You may yourself have this knowledge. You may also know that Piaget posited a sequence of cognitive stages through which humans, to varying degrees, pass. How many stages are there, according to Piaget? The received answer is: four; and in the fourth and final one, *formal operations*, neurobiologically normal humans can reason accurately and quickly over formulas expressed in the logical system known as first-order logic ($\mathcal{L}_I$).[3]

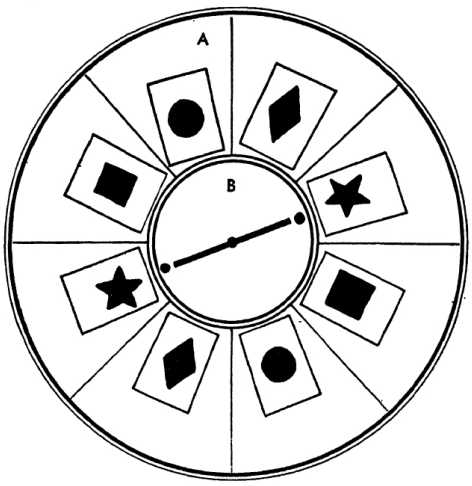

Fig. 1. Piaget's famous "rigged" rotating board to test for the development of Stage-3-or-better reasoning in children. The board, A, is divided into sectors of different colors and equal surfaces; opposite sectors match in color. B is a rotating disk with a metal rod spanning its diameter — but the catch is that the star cards have magnets buried under them (inside wax), so the alignment after spinning is invariably as shown here, no matter how the shapes are repositioned in the sectors (with matching shapes directly across from each other). This phenomenon is what subjects struggle to explain. Details can be found in [4].

Judging by the cognition taken by Piaget to be stage-three or stage-four (e.g., see Figure 1, which shows one

[1] Online, elegant, economical coverage can be found at http://plato.stanford.edu/entries/logic-classical/

[2] Many readers will know that Piaget's position long ago came under direct attack, by such thinkers as Wason and Johnson-Laird [5], [6]. In fact, unfortunately, for the most part people believe that this attack succeeded. Bringsjord doesn't agree in the least, but this isn't the place to visit the debate in question. Interested readers can consult [7], [8].

[3] Various other symbols are used, e.g., the more informative $\mathcal{L}_{\omega\omega}$.

of the many problems presented to subjects in [4]), the basic scheme is that an agent $\mathcal{A}$ receives a problem P (expressed as a visual scene accompanied by explanatory natural language), represents P in a formal language that is a superset of the language of $\mathcal{L}_I$, producing $[P]$, and then reasons over this representation (along with background knowledge Γ) using at least a combination of some of the proof theory of $\mathcal{L}_1$ and "psychological operators."[4] This reasoning allows the agent to obtain the solution $[S]$. To ease exposition, we shall ignore the heterodox operations that Piaget posits (see note 4) in favor of just standard proof theory, and we will moreover view $[P]$ as a triple (ϕ, C, Q), where ϕ is a (possibly complicated) formula in the language of $\mathcal{L}_I$, C is further information that provides context for the problem, and consists of a set of first-order formulas, and Q is a query asking for a proof of ϕ from $C \cup \Gamma$. So:

$$[P] = (\phi, C, Q = C \cup \Gamma \vdash \phi?)$$

For example, in the invisible magnetization problem shown in Figure 1, which requires stage-three reasoning in order to be cracked, the idea is to explain how it is that $\phi^{\star\star}$, i.e., that the rotation invariably stops with the two stars selected by the rod. Since Piaget is assuming the hypothetico-deductive method of explanation made famous by Popper [9], to provide an explanation is to rule out hypotheses until one arrives deductively at $\phi^{\star\star}$. In experiments involving child subjects, a number of incorrect (and sometimes silly) hypotheses are entertained — that the stars are heavier than the other shaped objects, that the colors of the sections make a difference, and so on. Piaget's analysis of those who discard mistaken hypotheses in favor of $\phi^{\star\star}$ is that they expect consequences of a given hypothesis to occur, note that these consequences fail to obtain, and then reason backwards by *modus tollens* to the falsity of the hypotheses. For example, it is key in the magnet experiments of Figure 1 that "for some spins of the disk, the rod will come to rest upon shapes other than the stars" is an expectation. When expectations fail, disjunctive syllogism allows $\phi^{\star\star}$ to be concluded. For our discussion of a sample functor over deductive systems as categories, it's important to note that while the hypotheses and context for the problem are naturally expressed using relation symbols, function symbols, and quantifiers from the language of $\mathcal{L}_I$, according to Piaget the final solution is produced by deduction in the propositional calculus.

III. From Piaget to Roboethics

What does all this have to do with roboethics? Well, for starters, notice that certain approaches to regulating the ethical decisions of lethal robots can be fairly viewed as aiming to engineer such robots by ensuring that they operate at Piaget's fourth stage. We believe this is true of both [10] and [11]. While in the first case an ethical code is to be expressed within some deontic/epistemic logic that subsumes classical logic,[5] and in the second there is no insistence upon using such more expressive logics, the bottom line is that in both cases there would seem to be a match with Piaget's fourth-stage: In both cases the basic idea is that robots work in a logical system, and their decisions are constrained by this work. In fact, it is probably not unfair to view an ethically relevant decision d by a robot to be correct if a formula in which d occurs can be proved from what is observed, and from background knowledge (which includes an ethical code or set of ethical rules, etc.) — so that a decision point becomes the solution of a problem with this now-familiar shape:

$$[P] = (\phi(d), C, Q = C \cup \Gamma \vdash \phi(d)?)$$

IV. The Intolerable Danger of Fourth-Stage Robots

In a sentence, the danger is simply that if a lethal agent is unable to engage in at least something close to sophisticated human-level ethical reasoning and decision-making, and instead can only operate at Piaget's fourth stage (as that operation is formalized herein), it is evident that that agent will, sooner or later, go horribly awry. That is, it will perform actions that morally wrong or fail to perform actions that are morally obligatory, and the consequences will include extensive harm to human beings.

The reason for such sad events will materialize is that a robot can flawlessly obey a "moral" code of conduct and still be catastrophically unethical. This is easy to prove: Imagine a code of conduct that recommends some action which, in the broader context, is positively immoral. For example, if human Jones carries a device which, if not eliminated, will (by his plan) see to the incineration of a metropolis, and a robot (e.g., an unmanned, autonomous UAV) is bound by a code of conduct not to destroy Jones because he happens to be a civilian, or be in a church, or at a cemetery ... the robot has just one shot to save the day, and this is it, it would be immoral not to eliminate Jones. (This of course just one from innumerable cases that can be easily devised.)

Unfortunately, the approach referred to in the previous section is designed to bind robots by fixed codes of conduct (e.g., rules of engagement covering warfighters). This approach may well get us all killed.

The approach that *won't* get us killed, and indeed the only viable path open to us if we want to survive, is to control robot behavior by operations over an ensemble of suitably stocked logical systems — operations from which suitable codes can be mechanically *derived* by robots on the fly. Once the code has been derived, it can be applied in a given set of circumstances.

V. But Then Why Piaget's Paradigm?

But if Piaget posits four stages, and deficient approaches to ethically correct robots already assume that such robots

[4] The psychological operators in question cannot always be found in standard proof theories. For example, Piaget held that the quartet I N R C of "transformations" were crucial to thought at the formal level. Each member of the quartet transforms formulas in certain ways. E.g., N is *inversion*, so that $N(p \vee q) = \neg p \wedge \neg q$; this seems to correspond to DeMorgan's Law. But R is *reciprocity*, so $R(p \vee q) = \neg p \vee \neg q$, and of course this isn't a valid inference in the proof theory for the propositional calculus or $\mathcal{L}_I$.

[5] Rapid but helpful overview of epistemic and deontic logic can be found in [12]. For more advanced work on computational epistemic logic see [13].

must operate at the fourth and final stage, what does the Piagetian paradigm have to offer those in search of ways to engineer ethically correct robots? The key fact is that Piaget actually posited stages *beyond* the fourth one — stages in which agents are able to operate over logical systems. For example, we know that logicians routinely create new logical systems (and often new components thereof that are of independent interest); this was something Piaget was aware of, and impressed by. But most people, even scholars in academia, are not aware of the fact that Piaget's scheme made room for cognition beyond the fourth stage. (Full references are forthcoming.)[6]

VI. Category Theory For Fifth-Stage Robots

Category theory is a remarkably useful formalism, as can be easily verified by turning to the list of spheres to which it has been productively applied — a list that ranges from attempts to supplant orthodox set theory-based foundations of mathematics with category theory [14], [15] to viewing functional programming languages as categories [3]. However, for the most part — and this is in itself remarkable — category theory has not energized AI or computational cognitive science, even when the kind of AI and computational cognitive science in question is logic-based.[7] We say this because there is a tradition of viewing logics or logical systems from a category-theoretic perspective. For example, Barwise [18] treats logics, from a model-theoretic viewpoint, as categories; and as some readers will recall, Lambek [19] treats proof calculi (or as he and others often refer to them, *deductive systems*) as categories. Piaget's approach certainly seems proof-theoretic/syntactic; accordingly, we provide now an example of stage-five category-theoretic reasoning from the standpoint of proof theory.

The example is based on two logical systems known to be directly used by Piaget, the propositional calculus $\mathcal{L}_{PC}$ and full first-order logic $\mathcal{L}_I$. Corresponding to both cases is a category, $\mathcal{C}_2$ and $\mathcal{C}_1$, respectively. In essentially a parallel to the treatment of deductive systems as categories,[8] we assume that *objects* in both cases are formulas generated by a recursive grammar operating over the standard logical symbols, and fixed non-logical alphabets. We let *arrows* be deductions, and we subscript and superscript δ to denote deductions. Thus,

$$\phi \xrightarrow{\delta} \psi$$

says that deduction δ goes from ϕ to ψ. Operations on arrows, for us, correspond to rules of inference; *composition* of deduction is straightforward.[9] The composition of δ and δ' is written

$$\delta;\delta'.$$

It should be obvious that the conditions required for a category are satisfied in this scheme.

Given $\mathcal{C}_1$ and $\mathcal{C}_2$, a particular functor from the former to the latter that is cognitively plausible from a Piagetian perspective is *. To formula ϕ_I of $\mathcal{C}_1$ is assigned $\phi_I^\star$, where of course $\phi_I^\star$ is a formula in the mere propositional calculus. This functor regiments what we noted to be happening in connection with the magnet mechanism above, viz., subjects are representing phenomena associated with the apparatus using relations and quantifiers, but then encoding this information in the propositional calculus. In addition, if in $\mathcal{C}_1$ there is

$$\phi_I \xrightarrow{\delta} \psi_I$$

no rules of inference for quantifier reasoning or identity are used in δ. For example, while $\forall x\psi \vee \forall x\psi'$ certainly partakes of the alphabet of $\mathcal{L}_I$ not shared with the alphabet of $\mathcal{L}_{PC}$, we have

$$(\forall x\psi \vee \forall x\psi')^\star = p \vee q$$

by straightforward algorithms with which many readers will be familiar.[10] Though we leave aside many details here, it should be possible for readers to see that certain desirable theorems are easy to establish, for example that the deduction in

$$(\phi_I)^\star \xrightarrow{\delta^\star} (\psi_I)^\star$$

is formally valid given that δ is.

It seems to us plausible that in the case of the magnet challenge, humans who successfully meet it essentially do a proof by cases, in which they rule out as unacceptable certain hypotheses for why the rod always stops at the stars. Assuming this is basically correct, it seems undeniable that though humans perceive all the relations that are in play (colors, shapes, and so on), and in some sense reason over them, something like the functor $\star$ is applied to more detailed reasoning in $\mathcal{L}_I$ to distill down to the core reasoning, which is expressible in $\mathcal{L}_{PC}$, and hence drops explicit reference to relations. The situation as we see it is summed up in Figure 2.

VII. Demonstrations

Demonstrations of actual robots operating on the basis of the approach described above will be available for the Roboethics Workshop at ICRA 09. Work toward such demonstrations is underway in the RAIR Lab. We seek robots able to succeed on Piaget's challenges (not only the "magnetic" problem of Figure 1, but others), on Piagetian challenges of our own design that catalyze post-stage-four reasoning and decision-making, and on microcosmic versions of the ethically charged situations that robots will

[6] This is as good a place to mention that we are in the process of exploring Kohlberg's work on moral/ethical development and reasoning.

[7] Bringsjord is as guilty as anyone, in light of the fact that even some very recent, comprehensive treatments of logicist AI and computational cognitive science are devoid of category-theoretic treatments. E.g., see [16], [17].

[8] We follow Lambek's original scheme [19], but we say 'essentially' here because, clearly, in Piaget's work, proof calculi for humans would not necessarily include the full machinery of standard ones for the propositional and predicate calculi, and moreover humans, according to Piaget, make use of idiosyncratic transformations that we would want to count as deductions (see note 4). Coverage of such details must wait for subsequent versions of the present paper.

[9] Arkoudas' NDL system, or a variant thereof, is a candidate for a calculus fitting the current context. It explicitly calls out deductions, which can be composed. For information, go to http://www.cag.lcs.mit.edu/∼kostas/dpls/ndl

[10] See the Truth-Functional Form Algorithm given in [1, Chapter 10].

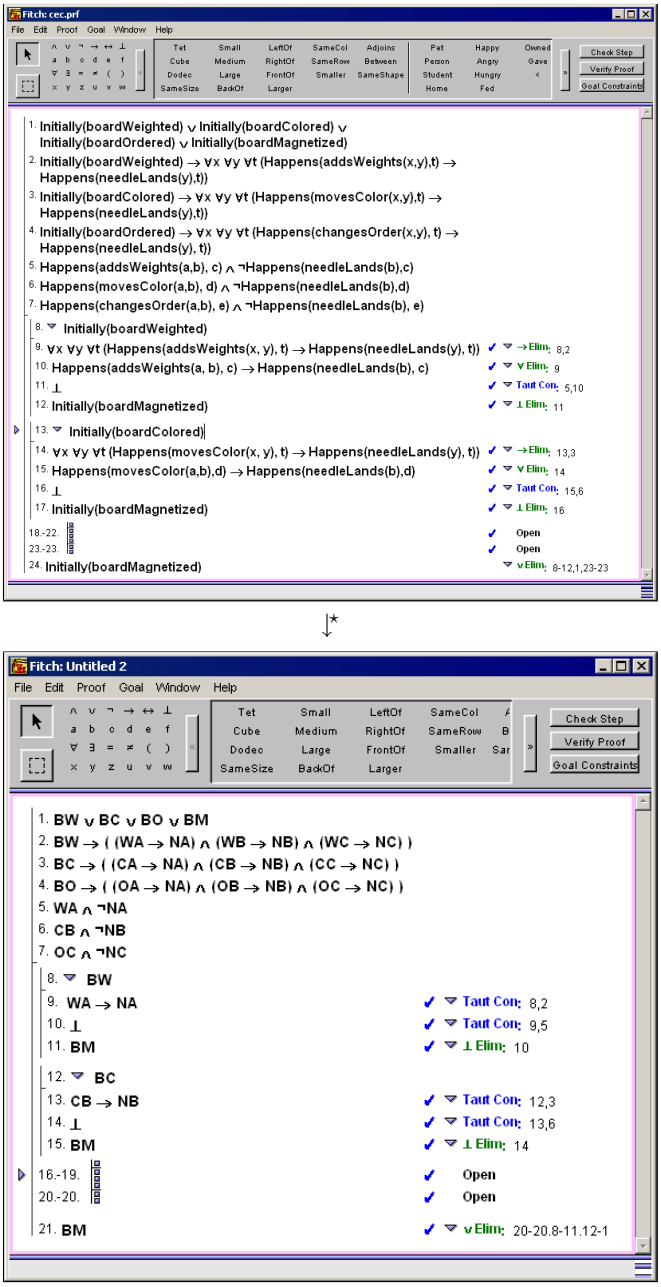

Fig. 2. This figure shows two proofs, one expressed in $\mathcal{L}_I$, the other in $\mathcal{L}_{PC}$. The first-order proof produces the conclusion that what causes the metal rod to invariably stop at the stars is that there are hidden magnets. The basic structure of this proof is proof by cases. Of the four disjuncts entertained as the possible source of the rod-star regularity, the right one is deduced when the others are eliminated. The functor $\star$ is shown here to indicate that the basic structure can be preserved in a proof couched exclusively in the propositional calculus.

see when deployed in warfare and counter-terrorism, where post-stage-four reasoning and decision-making is necessary for successfully handling these situations. The work here is connected to NSF-sponsored efforts on our part to extend CMU's Tekkotsu [20], [21] framework so that it includes operators that are central to our logicist approach to robotics, and specifically to roboethics — for example, operators for belief (**B**), knowledge (**K**), and obligation ($\bigcirc$ of standard deontic logic). The idea is that these operators would link to their counterparts in bona fide calculi for automated and semi-automated machine reasoning. One such calculus has already been designed and implemented: the *socio-cognitive calculus*; see [22]. This calculus includes the full event calculus.

Given that our initial experiments will make use of simple hand-eye robots recently acquired by the RAIR Lab from the Tekkotsu group at CMU, Figure 3, which shows one of these robots, sums up the situation (in connection with the magnet challenge). If sufficiently intricate manipulation cannot be achieved with the simple hand-eye robots, we will use the more powerful PERI, shown in Figure 4.

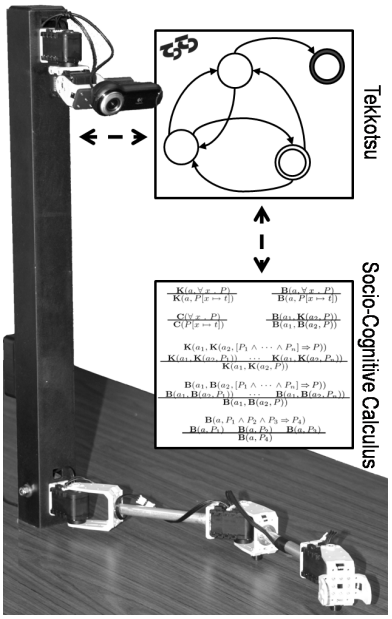

Fig. 3. The basic configuration for our initial implementations.

Fig. 4. The RAIR Lab's PERI

VIII. Future Research

Our title contains 'Piagetian Roboethics,' but the approach described in this short document can of course generalize to robotics *simpliciter*. This generalization will be pursued in the future. In fact, the direction described herein is the kernel of an approach to logicist AI and computational cognitive science, whether or not the agents involved are physical or non-physical. Therefore, in the future, the general concept of agents whose intelligence derive from reasoning and decision-making over logical systems (and their components) as categories will be pursued as well. Bringsjord believes that sophisticated human cognition, whether or not it is directed at ethics, exploits coordinated functors over many, many logical systems encoded as categories. These systems range from the propositional calculus, through description logics, to first-order logic, to temporal, epistemic, and deontic logics, and so on.

References

[1] J. Barwise and J. Etchemendy, *Language, Proof, and Logic*. New York, NY: Seven Bridges, 1999.

[2] H. D. Ebbinghaus, J. Flum, and W. Thomas, *Mathematical Logic (second edition)*. New York, NY: Springer-Verlag, 1994.

[3] M. Barr and C. Wells, *Category Theory for Computing Science*. Montréal, Canada: Les Publications CRM, 1999.

[4] B. Inhelder and J. Piaget, *The Growth of Logical Thinking from Childhood to Adolescence*. New York, NY: Basic Books, 1958.

[5] P. Wason, "Reasoning," in *New Horizons in Psychology*. Hammondsworth, UK: Penguin, 1966.

[6] P. Wason and P. Johnson-Laird, *Psychology of Reasoning: Structure and Content*. Cambridge, MA: Harvard University Press, 1972.

[7] S. Bringsjord, E. Bringsjord, and R. Noel, "In defense of logical minds," in *Proceedings of the 20th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum, 1998, pp. 173–178.

[8] K. Rinella, S. Bringsjord, and Y. Yang, "Efficacious logic instruction: People are not irremediably poor deductive reasoners," in *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society*, J. D. Moore and K. Stenning, Eds. Mahwah, NJ: Lawrence Erlbaum Associates, 2001, pp. 851–856.

[9] K. Popper, *The Logic of Scientific Discovery*. London, UK: Hutchinson, 1959.

[10] S. Bringsjord, K. Arkoudas, and P. Bello, "Toward a general logicist methodology for engineering ethically correct robots," *IEEE Intelligent Systems*, vol. 21, no. 4, pp. 38–44, 2006. [Online]. Available: http://kryten.mm.rpi.edu/bringsjord_inference_robot_ethics_preprint.pdf

[11] R. C. Arkin, "Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture – Part iii: Representational and architectural considerations," in *Proceedings of Technology in Wartime Conference*, Palo Alto, CA, January 2008, this and many other papers on the topic are available at the url here given. [Online]. Available: http://www.cc.gatech.edu/ai/robot-lab/publications.html

[12] L. Goble, Ed., *The Blackwell Guide to Philosophical Logic*. Oxford, UK: Blackwell Publishing, 2001.

[13] K. Arkoudas and S. Bringsjord, "Metareasoning for multi-agent epistemic logics," in *Fifth International Conference on Computational Logic In Multi-Agent Systems (CLIMA 2004)*, ser. Lecture Notes in Artificial Intelligence (LNAI). New York: Springer-Verlag, 2005, vol. 3487, pp. 111–125. [Online]. Available: http://kryten.mm.rpi.edu/arkoudas.bringsjord.clima.crc.pdf

[14] J.-P. Marquis, "Category theory and the foundations of mathematics," *Synthese*, vol. 103, pp. 421–447, 1995.

[15] F. W. Lawvere, "An elementary theory of the category of sets," *Proceedings of the National Academy of Science of the USA*, vol. 52, pp. 1506–1511, 2000.

[16] S. Bringsjord, "Declarative/logic-based cognitive modeling," in *The Handbook of Computational Psychology*, R. Sun, Ed. Cambridge, UK: Cambridge University Press, 2008, pp. 127–169. [Online]. Available: http://kryten.mm.rpi.edu/sb_lccm_ab-toc_031607.pdf

[17] ——, "The logicist manifesto: At long last let logic-based AI become a field unto itself," *Journal of Applied Logic*, vol. 6, no. 4, pp. 502–525, 2008. [Online]. Available: http://kryten.mm.rpi.edu/SB_LAI_Manifesto_091808.pdf

[18] J. Barwise, "Axioms for abstract model theory," *Annals of Mathematical Logic*, vol. 7, pp. 221–265, 1974.

[19] J. Lambek, "Deductive systems and categories i. Syntactic calculus and residuated categories," *Mathematical Systems Theory*, vol. 2, pp. 287–318, 1968.

[20] D. Touretzky, N. Halelamien, E. Tira-Thompson, J. Wales, and K. Usui, "Dual-coding representations for robot vision in Tekkotsu," *Autonomous Robots*, vol. 22, no. 4, pp. 425–435, 2007.

[21] D. S. Touretzky and E. J. Tira-Thompson, "Tekkotsu: A framework for AIBO cognitive robotics," in *Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI-05)*. Menlo Park, CA: AAAI Press, 2005.

[22] K. Arkoudas and S. Bringsjord, "Toward Formalizing Common-Sense Psychology: An Analysis of the False-Belief Task," in *Proceedings of the Tenth Pacific Rim International Conference on Artificial Intelligence (PRICAI 2008)*, ser. Lecture Notes in Artificial Intelligence (LNAI), T.-B. Ho and Z.-H. Zhou, Eds., no. 5351. Springer-Verlag, 2008, pp. 17–29. [Online]. Available: http://kryten.mm.rpi.edu/CognitiveCalculus092808.pdf