Cowley, M. (2006). The relevance of intent to human-android strategic interaction and artificial consciousness. *Proceedings, 15th International Conference on Robot-Human Interaction, IEEE*, 580-585.

The relevance of intent to human-android strategic interaction and artificial consciousness

M. Cowley
University of Southampton, UK

***Abstract*—A classic problem for artificial intelligence is to build a machine that imitates human behavior well enough to convince those who are interacting with it that it is another human being [1]. One approach to this problem focuses on building machines that imitate internal psychological facets of human interaction, such as artificially intelligent agents that play grandmaster chess [2]. Another approach focuses on building machines that imitate external psychological facets by building androids [3]. The disparity between these approaches reflects a problem with both: Artificial intelligence abstracts mentality from embodiment, while android science abstracts embodiment from mentality. This problem needs to be solved, if a sentient artificial entity that is indistinguishable from a human being, is to be constructed. One solution is to examine a fundamental human ability and context in which both the construction of internal cognitive models and an appropriate external social response are essential. This paper considers how reasoning with intent in the context of human vs. android strategic interaction may offer a psychological benchmark with which to evaluate the human-likeness of android strategic responses. Understanding how people reason with intent may offer a theoretical context in which to bridge the gap between the construction of sentient internal and external artificial agents.**

I. Introduction

STANLEY Kubrick's 2001: A Space Odyssey offers a tantalizing glimpse of what it might be like for humans to interact with a conscious artificial agent. The supercomputer 'HAL' (i.e., Heuristic Algorithm) is a built-in computerized member of a team of astronauts aboard a spacecraft on mission to Jupiter. As the mission progresses we first see HAL, a circular red interface, interacting with the astronaut Frank when they play a game of chess. Even though Frank is easily beaten, HAL subsequently compliments him on his play and consistently interacts with Frank in a complimentary and friendly manner. But unknown to Frank, HAL competes with his fellow astronauts and is secretly planning a strategy to kill off the astronauts one by one. While the consequences of this strategic behavior are unfortunate, we can say that HAL is exhibiting an activity fundamental to human consciousness [4], that is, HAL is reasoning with intent [5].

In this paper an examination of how androids may be perceived as reasoning with intent by their human counterparts will be examined. This conceptual examination will take place in the strategic context of chess playing. It is anticipated that future experimental work addressing the conceptual principles of intent, provided in this paper, may be one of the first important steps to corroborating the existence of an artificially constructed consciousness [1].

Thus, if agents appear to be able to consider alternative courses of action and to work toward an intended outcome by choosing an alternative based on anticipated consequences, they will have the same kind of consciousness that humans have [6]. The agent of specific interest in this paper are androids because they have been defined as artificial robotic agents that look and act like a human who aims to maintain human-like relationships with people [7]. In order to achieve this human-like interaction the android should be conscious of its own intentions, and to have the ability not only to conceal those intentions, but to be able to anticipate the intent of its human counterparts in the 'mind-reading' sense explicated by the Turing Test [1].

But while the original Turing Test assumes that there is 'little point in trying to make a thinking machine more human by dressing it up in… artificial flesh' [1, p.434], this paper proposes that this is precisely what now needs to be researched, and outlines a potential experimental program in which future studies may begin to address this question. The forward-looking idea of this paper is to examine one psychological benchmark related to the notion of embodied consciousness [33], such as the principle of intent in human reasoning [20], which may aid the future construction of artificially conscious agents.

One problem that occurs is which sort consciousness relates better to the psychological benchmark of reasoning with intent: sentience or sapience [34]? Sentience is defined as a capacity for a consciousness that has the ability to 'feel' [10]. The term sentience was chosen rather than sapience, a consciousness that knows, [10] for the following two reasons. The first reason why sentience is useful as a term to denote embodied consciousness is that artificial intelligence has tended to pursue the creation of machines with internal structures that know rather than feel [2]. The problem with machines that know is that they can appear to use knowledge in a human-like way, but do not necessarily emulate the human thinking process [35-36]. Human thinking may neither be separate from a sense of feeling, nor the corresponding physiological substrates associated with that

M. Cowley is with the Applied Social Sciences team, School of Social Sciences, University of Southampton UK, SO17 1BG (phone: +44 (0)23 80597451; e-mail: M.Cowley@soton.ac.uk).

sense of feeling [21], such as the sensitivity to future desirable or undesirable consequences when planning a course of action [22]-[31].

Consider the construction of grandmaster level chess playing programs such as Deep Blue who defeated the world chess champion Garry Kasparov [2]. While this result was a great achievement for artificial intelligence research, to conclude that Deep Blue's thinking was conscious in the human sense would be a mistake. This emulation of human-like superior cognitive performance is disembodied in the sense that the output from human and machine are at similarly high levels, but the processes by which human and machine produce this output are dissimilar. Deep Blue considered 90 billion moves at each turn, at a rate of 9 billion per second [8]. Computer chess programs still cannot perform to world champion standard without using extensive search [36]. The processes of grandmaster level chess playing programs are unlike those processes that human world champion chess players use [9]. When the output from internal symbol manipulation appears to show that the entity understands those symbols, we cannot conclude that the entity processes those symbols in a human-like way [10].

The second reason why sentience is a useful term is because a branch of robotics research, recently coined *android research* [13], has tended to construct machines with the external appearance and actions of a human, such as Ishiguro's android who can read the news [3]. But human-like external appearance does not mean that an android possesses properties that are intrinsic to knowing like a human [11]. To date android science tends not to construct androids which embody a human-like mentality [3].

II. ANDROIDS AND THEIR HUMAN COUNTERPARTS

Android science presently considers its primary problem to be the external design of human-like interactive robots. However android scientists have encountered a problem known as *the uncanny valley* [3]. In general it is accepted that as an entity approaches realistic human-likeness, from an initial cartoon likeness towards absolute human-likeness, there is a point at which the entity achieves a human-likeness that is perceived as 'eerie' or uncanny [3]. Typically, androids fall victim to this uncanny valley effect because they are perceived to be human-like but lacking something [3]-[11]. Bypassing the uncanny valley is essential for android science if we are to embrace a future in which social interaction with androids emulates human-like social interaction. But what is it that androids lack?

The uncanny valley has tended to be explained by external physical design problems such as the degree to which the android emulates human facial expression and is aesthetically pleasing [3]-[13]. On the one hand, consider how it could be difficult for the astronauts in Stanley Kubrick's 2001: A Space Odyssey to accept the internally contained HAL to be human-like. Even though HAL expresses linguistic terms and tone to indicate feeling, it may be difficult for human perception to interpret his feelings or believe they may be genuine, because he has no face or body to express them with [37-38]. On the other hand bodies that behave like humans while lacking mentality, or sentience, may resemble zombies rather than humans [7]. The lack of perceived sentience may also explain why people perceive a given android as uncanny or zombie-like, perhaps in the sense of reminding people of death [7].

By artificially embodying intent that is connected to the anticipation of consequences, forward-looking research could design a thinking machine to emulate human thinking processes with similar physiological construction, albeit with different materials [11]. An agent, for example an android who is capable of intending, may be capable of the sort of deliberative action [39], goal construction [31], and active existence akin to a higher level of human-like consciousness than a consciousness that simply perceives [40]. The suggestion of this paper then is that a fundamental problem of the uncanny valley may not only be the result of an external design problem, but also the result of an absence of perceived human-like sentience.

As a case in point consider an android that would typically be perceived as uncanny but displays the human-like ability to anticipate what its human counterpart is thinking, such as anticipating what move the person is thinking about playing in a game of chess. Even though the external appearance of the android is uncanny, subsequent interaction may lead its human counterpart to believe that it is sentient in the mind-reading sense outlined by the Turing Test [1]. The ability to reason with intent may facilitate perception of the android as human enough to bypass the uncanny valley. An illustrative example of this bypassing of the uncanny valley for people occurs when Ezmerelda is able to see beyond the Hunchback of Notre-Dame's uncanny appearance to the humanity that lies beneath [14].

Thus, the plan of the remainder of this paper will address two problems. The first problem, how artificial intelligence research tends to abstract mentality from embodiment and android science research abstracts embodiment from mentality, is addressed through the embodiment of intent [20]. It is asked whether an entity that resembles a human, such as an android, can bypass the uncanny valley effect? To address this question a future experimental program to examine the connection between internal cognitive models and subsequent social interactions in the strategic context of chess will be suggested. Second, the problem of reasoning with intent will be addressed. The main psychological theories of reasoning are outlined and the theoretical case for the inclusion of the principle of intent is made. The two questions about intent that will be asked are 1) how may an understanding of a principle of intent subsequently affect external actions in social interaction, such as strategic interaction between an android and a human counterpart? And 2) how might external actions carried out by an android, which are perceived to be sensitive to the intent of its human counterpart, affect human-android strategic interaction and the perception of the android's human-likeness? The connection is made between psychological theories of reasoning and the psychological benchmarks of the Turing Test through the principle of intent. That is, how the representation of an internal cognitive model may relate to sentience in the form of reasoning with intent [15], and how this intent may be a form of representation in an embodied consciousness [10].

III. INTENT, EMBODIMENT AND SENTIENT ARTIFICIAL AGENTS

To begin to understand how to bridge the gap between internal cognitive models and external social responses, it is necessary to first outline the theoretical threads that may connect intent, embodiment and sentient artificial agents.

Effective social interaction requires relating to the world in a way that reflects reality [16]. One must relate one's subjective internal mental states to the external world by directing ones mind to objects, states of affairs and people. The general term used to reflect this directed relationship is intentionality [12]. This intentional state is often assumed to consist of a representative context in a psychological mode [12]. In other words intent may be represented internally in the form of cognitive models [20]. But while major psychological theories of cognition focus on how people reason with internally represented mental models [18], they do not focus on intent as a principle in driving the representation of these mental models when people reason [17]-[19].

Cognitive experimentalists and psychological theories of reasoning tend to ignore the body in their theorising about the content and structure of thought [17]-[20]. Theories of cognition and cognitive theories of the psychology of reasoning tend to abstract the mind and its reasoning from the body with little exception [19].

As a result the inclusion of an android test-bed in cognitive science offers a powerful new apparatus with which to address questions that have been traditionally difficult for cognitive experimentalists [18], such as the problem of abstracting the mind from the body [21]. The principles of embodied cognition, such constructing a sentient cognitive agent, who has an understanding of other persons as intentional agents, tends to be defined by embodied synthetic or neurological substrates [5]. Android science may offer a future where such synthetic substrates can be pinned down for the development of human-like androids. For example, the development of experimental work on perceived 'embodied' cognitive representation in the context of androids competing with human counterparts in a game of chess; the classical drosophila for artificial intelligence research. This approach may offer hints about how to create a future embodied artificial agent that appears to be sentient, in that they are capable of understanding human mental states such as intention.

To address the issue of intent this paper will outline the main cognitive theories of how people reason, and how this principle of intent may be an important omission from contemporary theoretical frameworks.

IV. THE PRINCIPLE OF INTENT AND PSYCHOLOGICAL THEORIES OF REASONING

Contemporary psychological theories of how people reason tend to minimize the attention paid to the role that a person's intent may have in their thinking, whether they are planning an action, or reasoning with evidence towards a desired or undesired result [18-20]. Yet early philosophers tended to view every mental state as possessing the feature of intentionality [41]. For example, the word 'intent' or 'intention' tends to be synonymous with other nouns such as plan, purpose, objective, aim, target, etc. [24]. Each of these meanings can be understood as corresponding to an aim for a preconceived outcome [25]. For example, HAL's preconceived outcome was that the astronauts should be eliminated, and his aim or intention was to kill them in order to achieve that preconceived outcome. There are poignant everyday real world examples too. For example, juries must ascertain the existence of criminal intent and responsibility from presented evidence [26].

Consider the legal scenario in which Mr. X plans to kill Mr. Y who lives on the other side of town [27]. Mr. X gets into his car and places the weapon with which he plans to kill Mr. Y in the glove compartment. But unknown to Mr. X, Mr. Y is about to jog past his driveway. As Mr. X backs his car out of the driveway he does not see Mr. Y and runs him over killing him on the spot. A jury cannot conclude that Mr. X intentionally killed Mr. Y in this scenario [27-28]. People can understand that the actual outcome does not match Mr. X's preconceived outcome, in that Mr. X's intended method of killing Mr. Y does not match the actual killing method which was in fact unintentional [29].

It is possible that people can ascertain intentional and unintentional actions because they represent the actor's intention *in addition to* true states of affairs in the external world. For example, people may represent the possibility that Mr. X will drive to the other side of town to kill Mr. Y with a gun, AND they may represent the possibility that Mr. X backs out of his driveway and kills Mr. Y. Perhaps because people can represent both the unrealized intentional possibility and the realized actual outcome they can conclude that the killing of Mr. Y was unintentional. Neither possibility is consistent with the premise that Mr. X intended to drive across town to Mr. Y's house and actually killed him in this manner. A question that may be asked in the future is whether an android given such premises would conclude that Mr. X's killing of Mr. Y was unintentional. In other words would the android attend to external realities that actually happened, such as Mr. X's actual killing of Mr. Y by running him over, more readily than internal realities that did not actually happen, such as Mr. X's intention to drive across town to kill Mr. Y with a gun.

How might psychological theories of reasoning possibly accommodate the principle of intent? One possibility is that intent is one of the driving principles in what people internally represent in their thinking. That is, people may be able to represent both what they themselves intend and what they hypothesis that another may intend. There are currently three main cognitive theories of reasoning but we will focus on the most influential one—the mental model theory [18]. The mental model theory posits that people construct internal mental models representing possibilities corresponding to possible states-of-affairs in the world [18]. Consider what people may represent in the proceedings of the following criminal trial: A man smoked a cigar and was killed because an explosive was hidden inside it. The police find that the cigars in the man's cigar box have been skilfully rewrapped with explosive hidden inside them. Several strands of long hair are found underneath the cigars. The man's wife Martha has been accused of the murder [cited in 42].

> *Premises*
> If Martha's hair is in the box then she is the murderer…
> Martha's hair is in the box …
> *Conclusion*
> Martha is the murderer

We can see from this example that people may represent what they think is true [18], and according to the mental model theory people tend to represent models of what is true, *or what they believe to be true*, and not what is false in accordance with the **principle of truth**. But people may construct alternative models that may not be consistent with this logic, such as Inspector Wolfe who has a hunch that Martha is innocent [42]. He thinks of the alternative possibility in which the murderer has intended to frame Martha. After all, the skill required to neatly rewrap cigars with explosive hidden inside does not fit well with the carelessness of leaving one's hair alongside them. Inspector Wolfe has constructed a further possibility in which the conclusion that Martha is the murderer is invalid. However, in the end we discover that Martha was indeed the murderer and she intentionally placed the hairs alongside the cigars with explosives to make it look like she had been framed. The solution to this problem indicates that Martha was the murderer but the additional model that represents this possibility accommodates her intent.

Early research has shown that people remember intended actions that they did not carry out better than actions they did carry out [30]. The explanation was that intended actions require a latent activation in memory because they have yet to be carried out. Likewise, research into psychological disorders has shown that the inability to understand other minds and intentions is a primary cognitive feature of autism spectrum conditions [4], and the major debilitating feature is a chronic inability to engage in social interaction. Likewise an android who does not understand others' minds and intentions may be unable to socially interact.

For the sake of argument there will be a tentative assumption that a **principle of intent** may drive the construction of internal mental models, *or the most relevant mental model*, with which people reason in order to take part in social interaction, such as strategic interaction with an opponent [31]. From the above examples it may be considered necessary to understand more about how people may reason with intent in order to construct an artificial agent who can be perceived to reason with intent effectively.

The following section of this paper attempts to bridge the gap between internal cognitive models of intent and external social behaviors by focusing on this principle of intent as a psychological benchmark of human-like robotic agents in an android test-bed.

V. The Principle of Intent as a Psychological Benchmark for Android Science

The first step in being able to reason towards an understanding of other people's intentions is to have a concept that other people have a mind which is separate from physical reality [4-5]. Imagine we have created an android child and we ask her to listen to a story about two child characters. The first child called Sally is having a mental experience, for example 'thinking about a dog'. The second child called Molly is having a physical experience, for example 'holding a dog'. The experimenter then asks the android child which character can stroke the dog [32]? Children of 3-4 years of age can grasp the distinction between mental events and physical reality and they usually decide that Molly can stroke the dog, whereas children who do not understand that other people have mental states separate from reality, such as autistic spectrum children, tend not to be successful on this task [4].

But are errors of mind to mind interaction considered 'uncanny'? In other words, are such errors barriers to bypassing the uncanny valley effect in android science? Imagine our android child is expected to interact with other children perhaps by playing a game of strategy such as chess. The android child must understand the difference between a mental event and a physical reality in order to strategically interact with an opponent, for example to conceal one's intended plans [31]. For example, we can set an uncanny context, so that before the game begins the android child's human-child counterpart says that she is thinking about her pet dog, and the android child then asks if they can stroke the dog. The human child might become quite confused, or find the android child's response somewhat uncanny and play accordingly.

Social interaction requires an ability to think about a full range of mental states that are not externally visible to an android including beliefs, desires, emotions and intentions [4]. This ability is considered a fundamental human quality [5], and androids need to have or be perceived to have the ability to understand that their human counter-parts have minds [32]. If the android child does not display mentality awareness then it may not be likely that the human child will continue to play with it as an equal counterpart (implying no game of chess). Perhaps the child would find the android intriguing rather than uncanny, and may even seek to trick it in some way, for example, to tell the android to tidy away some toys. Either way, the social interaction may not achieve the human-like interaction and the maintenance of relationships with humans that android science aims for.

If android science pursues the route whereby androids are designed to respond to external stimuli [11-13], and understand mental events as external events, future androids may display a defect this paper will call **mind-blindness** [4-5], where there is an unlimited potential for social interaction errors due to the inability to discern the intent of social counterparts. In short, to pursue embodiment design without paying attention to the design of social interaction mechanisms that imitate human mentality (such being able to reason with intent), the uncanny valley problem may not be bypassed.

There are three starting points that may help to solve this problem. Firstly, the introduction of a human-android experimental paradigm in which the android is designed to respond to the human in a way that reflects the presence of an internal mentality. For example, an android child is given a short time interval for interaction with a human child in which an experimental psychologist administers some tests. The psychologist may ask both the android child and the human child to sit with one another and to answer questions concerning the mental-physical distinction as in the Sally-Molly dog story we saw earlier. The experiment could measure the human child's reaction when the android child consistently gives the incorrect or uncanny answers such as asking to stroke the mental dog. If the human child is less willing to engage in a game with this mind-blind android than an android child that is not mind-blind, we may conclude that effective androids should exhibit a sentient mentality in addition to an aesthetically pleasing design.

Second, how may children decide to subsequently engage with the following types of child androids: 1) an aesthetically pleasing android who is not mind-blind, 2) an aesthetically pleasing android who is mind-blind, 3) a non-aesthetically pleasing android who is not mind-blind, and 4) a non-aesthetically pleasing android who is also mind-blind? If the ability to discern intent is a factor that helps androids maintain human-like relationships, then the desire for subsequent engagement would be more likely with an aesthetically pleasing android who is not mind-blind, than with an aesthetically pleasing android who is mind-blind. Likewise, the desire for subsequent engagement with a non-aesthetically pleasing android who is not mind-blind would be more likely than with a non-aesthetically pleasing android who is mind-blind.

Third, an experimental paradigm in which adult humans and androids interact in a strategic context, such as chess playing may provide a context for understanding reasoning with intent in human-android strategic interaction. The chess domain may be a more traditional way to address the problem of android consciousness in social interaction [2]-[31]. E.g., the chess player is instructed to interact with the android before they start to play 'to practice interacting with an android'. But unknown to the player, the standard questions the experimenter gives to practice when the android is present, are an opportunity for an intentionality awareness test (e.g., a variation on the Sally-Molly dog problem). In one condition the android will answer correctly and display mentality awareness, and in another condition the android will answer incorrectly and so not display mentality awareness. The human player and the android are then presented with a chess position; it is the human player's turn to move. The player is told that the android has not seen this position before and that the android has been programmed to play chess in a human-like way. The most likely courses of action available to the chess player have previously been worked out by the experimenter; the experimenter is aware of all the most likely moves the player can choose The position is set up such that there is at least one enticing possibility for the human player which is a win, should the android not be aware of a defending move placed some moves ahead that is difficult to spot. The android will always know to play at the appropriate time.

If the player is convinced that the android is aware of intent based on their initial interaction, he may assume that the android will be able to anticipate what he/she intends to play, and not choose the enticing move. If the player is not convinced that the android is aware of intent based on their initial interaction, he may choose the enticing move in hope that the android will not spot what he intends to do [31]. In other words the human chess player's strategic choice may be affected by his/her assumption of the android's ability to represent an opponent's intent [16]-[31].

Imagine the human chess player who chooses the enticing move in the hope that the mind-blind android will not spot the accurate counter move to defend against the human player's attack. The android effectively defends ands says 'aha I thought you might but you won't trick me!' How might such a strategic interaction affect the human player's perception of the human-likeness of the android? For example, if human chess players compared the android with a chess playing interface such as Fritz 9 who says the same words in the same tone etc., would they rate the android as more human-like because it appears to have an embodied intent? Perhaps the human will rate the android as more human-like regardless of its uncanny appearance or initial failure at the Sally-Molly dog test.

In conclusion, a case was made for the idea that reasoning with intent is a key psychological benchmark with which to evaluate human-like artificial agents. It was suggested that this psychological benchmark should be embodied within an artificial agent, such as an android. The notion of an embodied intent may bridge the gap between disparate AI approaches that focus on internal or externally constructed artificial agents. The revolutionary idea is that research with androids that appear to be able to discern human intent, may help narrow the gap between the construction of robotic agents who are perceived to possess human-like consciousness, and future robotic agents that may actually possess artificial consciousness.

Acknowledgment

The author gratefully acknowledges the encouragement and work of Karl MacDorman, and two anonymous reviewers for their helpful insights.

References

[1] Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, vol. LIX, no. 236, pp. 433-451.

[2] Hsu, F., M. S., Campbell, & Hoane, A. J. (1995). Deep Blue system overview. In the *Proceedings of the Ninth International Conference on Supercomputing*. USA: ACM Press, pp. 240-244.

[3] MacDorman, K. F., Minato, T., Shimada, M., Itakua, S., Cowley, S., & Ishiguro, H. (2005). Assessing human-likeness by eye contact in an android test-bed. In B. Bara, L. Barsalou, & M. Bucciarelli., the *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*. Mahwah: Lawrence Erlbaum, pp. 1373-1378.

[4] Baron-Cohen, S. (2001). Theory of mind in normal development and autism. *Prisme*, vol. 34, pp. 174-183.

[5] Baron-Cohen, S. (2002). The intentional stance: Developmental and neurocognitive perspectives. In A. Brook, & D. Ross (Eds.), *Daniel Dennett*. Cambridge: Cambridge University Press, pp. 83-116.

[6] Nilsson, N. (1980). *Principles of Artificial Intelligence*. USA, San Francisco: Morgan Kaufmann.

[7] MacDorman, K. F. (2006). The Uncanny Advantage of using Androids in Cognitive Research. To appear in *Interaction Studies*.

[8] Eysenck, M. W., & Keane, M. T. (2005). *Cognitive Psychology: A student's handbook*. Hove, UK: Psychology Press.

[9] Chabris, C. F., & Hearst, E. S. (2003). Visualisation, pattern recognition and forward search: Effects of playing speed and sight of the position on grandmaster chess. *Cognitive Science*, vol. 27, pp. 637-648.

[10] Searle, J. (1983). *Intentionality: An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press.

[11] MacDorman, K. F. (2004). Extending the medium hypothesis: The Dennett-Mangun controversy and beyond. *Journal of Mind and Behaviour*, vol. 25(3), pp. 237-257.

[12] Matsue, D., Minato, T., MacDorman, K. F., & Ishiguro, H. (2005). Generating natural motion in an android by mapping human motion. *Proceedings the IEEE/RSJ International Conference on Intelligent Robots and Systems*. Edmonton, Canada, 2 -6 August, 2005.

[13] MacDorman, K. F., & Ishiguro, H. (2005). Toward social mechanisms of android science: A CogSci 2005 Workshop. In *Interaction Studies*, vol.7(2).http://www/macdorman.com/kfm/writings/pubs/MacDorman 2005TowardSocMechIntStudies

[14] Hugo, V. (2004). *The Hunchback of Notre-Dame*. London: Penguin

[15] Johnson-Laird, P. N. (1983). *Mental Models*. Cambridge: Cambridge University Press..

[16] Cowley, M. (2006). *The Role of Falsification in Hypothesis Testing*. PhD Thesis: Trinity College Dublin.

[17] Craik, K. (1943). *The Nature of Explanation*. Cambridge: Cambridge University Press.

[18] Johnson-Laird, P. N. (1999). Deductive reasoning. *Annual Review of Psychology*, vol. 50, pp. 109-135

[19] Evans, J. St. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences*, vol. 7, pp. 454-459.

[20] Bara, B. (2006). Embodiment of Intentions. Paper presented at the *International Meeting on Mental Models and Reasoning*, Trinity College Dublin.

[21] Damasio, A. R. (1994). *Descartes Error: Emotion, reason and the human brain*. London: Papermac.

[22] Bechara, A. , Damasio, A. R., Damasio, H., & Anderson, S. W.(1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, 50, 7-15.

[23] Bechara, A. , Damasio, H., Damasio, A. R., & Lee, G. P. (1999). Different contributions to the human amygdala and ventromedial prefrontal cortex to decision making. *Journal of Neuroscience*, 19, 5437-5481.

[24] Oxford University Press (2002). *Little Oxford Thesaurus*. Oxford: Oxford University Press.

[25] OED (2002) *Oxford English Dictionary*. Oxford: Oxford University Press, 2002.

[26] Roberts, P., & Zuckerman, A. (2004). *Criminal Evidence*. Oxford: Oxford University Press.

[27] Harman, G. (1986). *Change in View*. Cambridge, MA: MIT Press.

[28] Murphy, P. (2005). *Murphy on Evidence*. Oxford: Oxford University Press.

[29] Jones, M. A. (2005). *Textbook on Torts*. Oxford: Oxford University Press.

[30] Vygotsky (1986). *Thought and Language*. London: MIT Press.

[31] Cowley, M., & Byrne, R. M. J. (2004). Chess Masters Hypothesis Testing. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, Chicago, 24-26 July, pp. 250-255.

[32] Wellman, H., & Estes, D. (1986). Early understanding of mental entities: A reexamination of childhood realism. *Child Development*, vol. 57, pp. 910-923.

[33] Lakoff, G., & Johnson, M. (1999). *Philosophy in the Flesh: The embodied mind and its challenge to Western thought*. New York: Basic books.

[34] Dennett, D. C. (2001). Are we explaining consciousness yet? *Cognition*, 79, 221-237.

[35] Gobet, F. (1997). Can Deep Blue make us happy? Reflections on human and artificial expertise. *AAAI-97 Workshop: Deep Blue vs Kasparov: The Significance for Artificial Intelligence*, p.20-23. AAAI press: Technical report WS-97-04.

[36] Gobet, F., de Voogt, A., & Retschitzki, J. (2004). *Moves in Mind: The Psychology of Board Games*. Hove, UK: Psychology Press.

[37] Ekman, P., & Friesen, W. (1976). *Pictures of facial affect*. Paolo Alto, CA: Consulting Psychology Press.

[38] Blakemore, S.-J., Winston, J., & Frith, U. (1998). How do we predict the consequences of our actions? A functional imaging study. *Neuropsychologia*, 36, 521-529.

[39] Lowe, E. J. (1996). *Subjects of experience*. Cambridge: Cambridge University Press.

[40] Crick, F. (1994). *The astonishing hypothesis: the scientific search for the soul*. London: Touchstone.

[41] Bartok, P. J. (2005). Brentano's Intentionality Thesis: Beyond the analytic and phenomenological readings. *Journal of the History of Philosophy, vol. 43(4)*, 437-460.

[42] Byrne, R. M. J.(2005). The Rational Imagination: How people create alternatives to reality. Cambridge: MIT Press.