Computationally Constrained Beliefs

Drew McDermott

Yale Computer Science Department

drew.mcdermott@yale.edu

J. of Consciousness Studies **20**(5–6), pp. 124–50

Abstract

People and intelligent computers, if there ever are any, will both have to believe certain things in order to be intelligent agents at all, or to be a particular sort of intelligent agent. I distinguish *implicit* beliefs that are inherent in the architecture of a natural or artificial agent, in the way it is "wired," from *explicit* beliefs that are encoded in a way that makes them easier to learn and to erase if proven mistaken. I introduce the term IFI, which stands for *irresistible framework intuition,* for an implicit belief that can come into conflict with an explicit one. IFIs are a key element of any theory of consciousness that explains qualia and other aspects of phenomenology as second-order beliefs about perception. Before I can survey the IFI landscape, I review evidence that the brains of humans, and presumably of other intelligent agents, consist of many specialized modules that (a) are capable of sharing a unified workspace on urgent occasions, and (b) jointly *model* themselves as a single agent. I also review previous work relevant to my subject. Then I explore several IFIs, starting with, "My future actions are free from the control of physical laws." Most of them are *universal*, in the sense that they will be shared by any intelligent

agent; the case must be argued for each IFI. When made explicit, IFIs may look dubious or counterproductive, but they really are irresistible, so we find ourselves in the odd position of oscillating between justified beliefs$_E$ and conflicting but irresistible beliefs$_I$. We cannot hope that some process of argumentation will resolve the conflict.

1 Introduction

This paper is an exploration of the notion of *belief whose truth is less important than its inescapability*, that is, the sort of thing that Friedrich Nietzsche might have had in mind when he observed that "untruth" might be "a condition of life" (Nietzsche, 1886, p. 216). An example is the belief that "Some possible futures are better than (preferable to, more satisfactory than) others." Some such beliefs are true, some are false, and some have truth values that are hard to judge.

A central topic will be the possibility of the simultaneous existence of stable contradictory beliefs in the same intelligent system. It will often become very confusing what kind of belief we are talking about. So I will sketch them up front and introduce a notation for distinguishing them.

For academics, the easiest-to-visualize sort of belief is an inscription in a *language of thought* (LOT) placed in a hypothetical location called the *belief box,* a caricature of a presumed actual locus or set of loci in the brain where active beliefs play a role in inference and decision (Schiffer, 1972). (To make a complete belief-desire psychology, just add a desire box, also containing LOT expressions.) Even if you are a connectionist, you must grant the existence of learned, conscious, erasable beliefs, stored somehow in neural nets. I will refer to this sort of belief, however it is implemented in people, as *belief*$_E$ ("e" for "explicit," "erasable").

We can contrast them with beliefs that emerge from the way a system is "wired." We know that the typical heterosexual male adult believes young women are interesting to look at because he looks at a lot of them, and, when made aware of this fact, keep right on doing it, with or without a blush. Another example, due to Dennett (1977) is a chess program that believes (erroneously) "it should get its queen out early." This claim is based on observation of many games in which that's what the program does. Nowhere is such a belief written down in a database of the program's beliefs, or even encoded in a neural net; it's just an indirect consequence of the program's position-evaluation procedure. I will use the term *belief*$_I$ for this sort of belief ("i" for "implicit," "inherent"). The same subscripts can be attached to other mentalistic words such as "desire" and "think" with similar intent.

IFIs are relevant to the study of consciousness because any theory that locates consciousness in the way an organism or other information-processing system models its own thought processes (such as those of Dennett (1991), McDermott (2001), or Metzinger (2003)) will locate qualia and other aspects of phenomenology as entities that exist because of beliefs$_I$ that they exist.

Belief$_I$ does not have to be innate or instinctive. But if a belief$_I$ is learned, it is learned in a "write-only" mode; and, in humans and other animals, at early life stages. A good example is ducklings' learning of who their mother is (Tinbergen, 1951). I will use the term *irresistible framework intuition* or *ifi* (pronounced "iffy") as a synonym for "belief$_I$," especially when the belief is unsupported by evidence, or in conflict with it. The word "irresistible" does not mean that there is absolutely no way to escape them. But overcoming more deep-rooted IFIs requires drastic measures, such as the use of Buddhist or other meditation techniques. These techniques do amount to dismantling oneself, or close to it. Nothing short of this would

seem capable of allowing practitioners to endure suicide by self-immolation with equanimity, for example.

Of course, normally the system of beliefs$_E$ is in harmony with the system of beliefs$_I$. One might believe in both ways, for instance, that root beer tastes better than ginger ale. But conflict is more intriguing than harmony, and that will be my focus. (In a context where this conflict is either absent or irrelevant, I will omit subscripts on words like "belief.") In addition, I am looking for IFIs of the broadest possible scope, so another focus of this paper is *universal IFIs,* those that are shared by all intelligent agents; and *global IFIs*, those shared by all humans.

Here is how the remainder of the paper is organized. In section 2.1 I review assumptions based on results in cognitive science about the organization of the brain/mind. In section 2.2 I look at previous relevant work. In section 3 I survey some IFIs that creatures like us must have. Section 3.1 catalogues universal IFIs concerned with freedom of decisions from causality ("free will"). Section 3.2 is about IFIs about perception and qualia. Section 3.3 draws IFIs from the work of Derek Parfit, showing that many of his counterintuitive conclusions are impossible for humans, or in some cases any intelligent agent, to actually believe. Finally, section 4 summarizes my conclusions.

2 Background

My goal is to explore the notion of irresistible framework intuitions (IFIs), that is, beliefs built into an intelligent agent that it is constrained to live with. But first we must address the question, What sort of computational structure do intelligent agents have?

2.1 Architectural Assumptions

It is fashionable nowadays to treat questions about morality, reason, and other human capabilities from the point of view of evolution by natural selection (Tooby and Cosmides, 1992, 2005; Dennett, 2003; McKay and Dennett, 2009). Here I take a different tack, and try to find constraints on intelligent agents imposed on them simply *because* they are intelligent agents. In other words, this really is a philosophy paper and not an evolutionary-psychology paper. I assume without much argument that to be an intelligent agent requires being a *situated computational* agent. What I mean by "situated" is simply that the agent behaves with respect to, that is, senses and operates upon, the real world (Steels and Brooks, 1995).

In what follows, when I speak of the "function" of a cognitive mechanism or trait, I intend to mean "what it does," and avoid any allusion to "what its *purpose* is." In particular, I don't mean to appeal to natural selection to account for a particular aspect of the way people are, even though I don't doubt that natural selection is responsible for the way they are. It would be nice to have detailed knowledge about how our traits fell into place, one by one, but we don't. However, we can still agree that the function of the heart is to pump blood in the sense that it does actually pump blood; and the function of the brain is to compute things in the sense that it does compute things.[1] (For a careful account of what "compute" means, see McDermott (2001).)

I assume that deriving useful actions from sensory inputs requires computation.

[1] One might counter that the function of the heart is to produce a certain amount of heat, because it does that, too. I have no problem with that; I am not trying to provide an analysis of what we normally mean by "function." I merely wish to call attention to things that hearts and brains do that are *interesting* for some reason.

So far we know of only one kind of intelligent agent, namely, human beings, where I take an *intelligent agent* to be one possessing language that is capable of imaginative projections of itself into the future in order to solve problems. (It's possible, but not important here, that other species possess these abilities to some extent.) But there seems to be growing certainty that AI will produce new breeds of intelligent systems, probably quite different in many respects from us (Kurzweil, 2005; Chalmers, 2010). (I use the word *projection* here in a technical sense to mean a model of a possible future. I reserve the word *prediction* to mean a best guess as to what will actually happen given what has happened so far.)

To date the products of AI labs, while impressive, exhibit what Kurzweil (2005) calls "narrow intelligence," the ability to perform one complex task at or above the level of human performance. Artificial agents can now control cars through realistic urban landscapes (Belfiore, 2007), and beat human champions at Jeopardy (Ferrucci et al., 2010). But each such agent, once pushed out of the "domain" it was designed to excel in, is incapable of even trying to compete.

The best face we can put on this narrowness is that AI, like other branches of cognitive science, still has work to do discovering the many specialized modules apparently required by intelligent systems, whose role is to solve well-defined problems that arise repeatedly, such as face recognition (Kanwisher et al., 1997) and spatial navigation (Gallistel, 1990).[2]

The brain can't, however, be a mere menagerie of special-purpose systems, if

[2]There is a group of people whose goal is to reform or redo AI so it is broad from the base up; they use the phrase "artificial general intelligence" (AGI) to describe this target. Although there is now a yearly conference held to present results of this kind, so far there have been no breakthroughs. See (Goertzel et al., 2009; Baum et al., 2010; Schmidhuber et al., 2011).

for no other reason than that its owner, being a single body, must allow at most one of them to control its behavior at any given time. Simple organisms can get by with a simple priority hierarchy that, for instance, makes sure that the *forage* behavior is suppressed when *flee from predator* is active. But at some point in our evolution, the brain began to do more than let one of its circuits take control; it recruited multiple circuits to solve the same problem. A rustle in the bushes catches our attention, and we turn our eyes in that direction. Now the ears and eyes are both at work on the problem *assess danger/opportunity in bushes.* They apparently pool their resources in a *global workspace* (Baars, 1988, 1997).

Specialized modules are reactive, whereas the "global" system takes a somewhat longer view, collecting information and assembling it into a *model* of the situation (Dennett, 2003). As Akins (1996) points out, one of the commitments that brainy creatures undertook was the "ontological project": keeping track of objects around them. Frogs don't care which fly they detect and eat; they probably don't care which pond they jump into to escape predators. But mammals do try to return to a particular place to sleep at night; birds do care which nest they bring food back to. Beginning with commitments to keep track of objects such as sleeping and nesting places, and conspecifics competing for them, evolution produced in the human species the ability to keep track of thousands of entities. But as soon as brains made this shift toward ontology the issue arose, *Who* is keeping track of these objects and their properties? What does it mean for "the brain" to represent something, as opposed to a specialized circuit within the brain?

We are now brought back to the question of the gap between the "narrow" accomplishments of AI so far and the expectations that there will exist "strong" artificial intelligent agents in the not too distant future, because the gaps in our

understanding of belief structures in the brain involve many of the same issues. Although there is a computational theory of knowledge representation (KR), it doesn't support optimistic expectations for efficient inference algorithms with expressive representations (Russell and Norvig, 2010). How does the brain do it? Neuroscience has yet to explain how to get from a bunch of neural nets, each maintaining a specialized, distributed code (Abbott and Sejnowski, 1999) to a ... what? A mega-net maintaining a general-purpose code?[3]

My tentative conclusion is that the following missing pieces must fall into place both for cognitive science to triumph *and* for AI to bring forth artificial intelligent agents:

1. We must learn more about the "language of thought" (LOT) used by the brain (Harman, 1973; Fodor, 1975). (But I think of this language as the medium of communication and computation rather than the medium of belief, at least the beliefs I am interested in; see below.)

2. There must be a computational theory of KR and inference using this language, and how these are embodied in neurons efficiently. This must include a theory of what Newell (1969) calls "weak methods," for the brain to fall back on when its specialized modules give up. The paradigmatic example of a weak method is *analogy,* which gets no respect for producing compelling conclusions, but which seems nonetheless to be ubiquitous in human thought (Lakoff and Johnson, 1980) and for which existing computational theories (Hofstadter, 1995; Falkenhainer et al., 1989; Forbus et al., 1994) fall short.

3. The purpose of language must be made clearer, in order to have a chance of de-

[3]See (McDermott, 2011a) for skeptical rumination on this topic.

> veloping computational theories of linguistic meaning. Existing theories place heavy emphasis on communication of beliefs and on logical-form ambiguities, which seems misplaced.

You can view it as a disappointment that cognitive science has yet to fill in these gaps, or as an achievement to have delineated them.

My intention is to be neutral between choice of algorithm, either based on combinatorial search using symbols, or based on number propagation and weight learning, as in artificial neural networks. It may seem that my insistence that we need an account of some kind of LOT puts in me in the previous camp.[4] But consider that even a staunch dynamicist such as Murray Shanahan (2010), who thinks even neural networks are at too high a level, and that we should be using the language of differential equations to describe brains, cannot help but lapse into linguistic expressions to express the messages traveling between brain regions. For someone in that camp, the urgent problem is to explain the LOT away, i.e., explain how to do without it and what takes its place. It is conjectured that synchronization between distant brain regions is a key factor (Dehaene and Naccache, 2001), but what information is conveyed during synchronization, and how, is still mostly a mystery.

So far I have focused on constraints on an agent derived from the fact that it is *an* agent, that is, a single entity. There are other properties due to the fact that each agent *thinks* of itself as a single entity existing through time. When we think of where we will be in the future we think of a particular entity, not a coalition of modules with possibly disparate motives.

[4]Which often is tagged with the rubric "GOFAI" — good old-fashioned AI (Haugeland, 1985). But if there is such a thing, the key change between the "old days" and now is the exponential increase in applications of statistics to AI problems.

Of course, when one speaks of how "we," or more precisely, "I" think of myself, it is hard not to think in circles. If an agent thinks of the future in terms of the adventures of a single agent, the attributes of that agent are the attributes it believes to be true of *itself*. But people are notoriously bad at knowing their true properties even in areas where objective statistics can be defined. The *locus classicus* for this view is (Nisbett and Wilson, 1977), which reviews evidence that people know little about the processes influencing their decisions. They are never really certain about whether they are deciding or being manipulated (Wegner, 2002). They also tend to overrate their abilities (McCormick et al., 1986). It is even argued that we know remarkably little about the events we report on introspectively (Dennett, 1991; Noë, 2002; Schwitzgebel, 2008).

If I am the person I identify with in my future projections, who or what is that really? Am I the animal owning the brain that does the thinking? Or the character the thinking seems to be about, that deeply thoughtful, caring, sensitive young man with an important message the world is dying to hear? Identifying the two seems unsatisfactory, but there seems to be no other choice. Contrast trying find the real entity with the best claim to be Santa Claus. For a Christian capitalist child who has been clued in, there is a choice between, on the one hand, concluding that Santa Claus *is* their parent(s), except for a few discrepancies such as what gender, what age, and even how many people Santa is supposed to be; or concluding that, alas, there really is no Santa Claus. But I can hardly conclude there is no I. The result is a system in which beliefs_I and beliefs_E about the self are often in conflict.

2.2 Previous Work

One of my main themes is the irresolvability of some contradictions between our explicit and implicit beliefs. Galen Strawson describes a similar clash (1986, p. 47n) of intuitions, one having to do with compatibility or incompatibility of free will and determinism:

> The conflict of intuition that produces this reversal or oscillation of views (the psychological explanation of which is a crucial part of a full account of the problem of freedom) is like a perpetual-motion machine. It promises to provide a source of energy that will keep the free will debate going for as long as human beings can think.

This is an apt description of the *kind* of conflict I have in mind. But Strawson, like many other philosophers (one can start with (Sidgwick, 1907; Smart, 1961; Strawson, 1962; Inwagen, 1983)) insists on linking free will, moral responsibility, and the experience of agenthood. I think the basic clash lies deeper, in what it *means* to make a conscious decision, not what it *feels like*. (See page 21, below.)

The only thing we're sure about if two belief systems disagree is that at least one of them is wrong. One of the first modern philosophers to call attention to the pervasiveness of error in our belief systems was J.L. Mackie (1977), who pointed out that almost everyone (except for a few sociopaths?) has an *erroneous* notion of right and wrong, erroneous because it conflicts with the fact of what I will call *scientific moral relativism,* the viewpoint from which value systems are objects of study, not frameworks one can be committed to. He also illustrates vividly our inability to free ourself even from errors we can clearly perceive.

Mackie's observation is that, if no value system can claim to be correct, then we all accept one or another system of systematic *falsehoods*. Mackie himself provides a perfect illustration of how difficult it is to escape the error he identified. Having announced that the entire ethics enterprise is riddled with absurdities, he goes on to advocate his own system of ethics.[5] But we should not throw stones at Mackie. I think it was a brilliant insight to see that, even though moral concepts are intrinsically nonrelative, and hence vacuous, applying to nothing, social groups have no trouble living by them for generations. In this paper, morality is not the main focus. Instead we ask, *Where else* do we find errors of the sort Mackie has identified? We discover we live in glass houses ourselves.

The most obvious issue for which there is a discrepancy between beliefs$_E$ and beliefs$_I$ is free will (Inwagen, 1983). P.F. Strawson's influential essay "Freedom and Resentment" (1962) is the modern source for the argument that our insistence on moral responsibility offers some pushback against any belief in determinism. (See also (Denyer, 1981).) The idea that we must live with multiple belief systems is broached by Van Inwagen (1983) but rejected. By nature philosophers would prefer almost any way of eliminating a contradiction to living with one; see (McDermott, 2011b) for discussion of why inherent beliefs are not easily changed, even in computer programs.

McKay and Dennett (2009) survey candidates for *adaptive misbeliefs,* false propositions that it might be adaptive (in the evolutionary sense) to believe. They clarify many of the issues involved in dealing with the structure of belief, but focus their lens as narrowly as possible, on beliefs that enhance their possessors' self-esteem.

[5] "It is almost as if he had first demonstrated that God does not exist and had then gone on to consider whether He is wise and loving" (Harman, 2000, p. 80).

My concern is with beliefs wired deeper and at a larger scale.

Tamar Szabo-Gendler (Gendler, 2008, 2009) has identified a rather different sort of mistake she calls *alief.* Aliefs are quasi-beliefs that we reject as inconsistent with our overall belief system, but that keep coming back. She cites many examples, including one's reaction to a glass walkway with a mile-deep drop underneath it; the point being that almost everyone reacts as though they[6] fear falling no matter how improbable they know this to be. The difference between the typical IFI and the typical alief is that the former is believed and the latter is not (instead being manifested by palpitations, disgust, or the like). Nonetheless, Gendler cites examples where aliefs are strong enough to defeat the beliefs that contradict them, as when an agoraphobic constructs a belief system that justifies staying home.

An important inspiration for my discussion of free will in section 3.1 was Dennett's work on this topic (Dennett, 1984). The idea that a person cannot analyze or understand themselves is discussed at length by (Ryle, 1949). But the issue goes back to Hume — of course! In the last half century, adaptation of these arguments for automata has naturally occurred. Karl Popper (1950a, 1950b) gives an ingenious argument for why, even in a deterministic universe, a predictor cannot predict itself (and two predictors cannot predict each other).

My proposals about self-modeling and qualia in section 3.2 are consistent with previous proposals by Lycan (1987,1997), Dennett (1991), McDermott (2001, ch. 3), Metzinger (2003), and Sloman and Chrisley (2003).

There is a large literature on what the goals of AIs will be if and when they come, and how much we should worry about those goals. Steve Omohundro (2008) believes that AIs will be hard to control and their utility functions tricky to design.

[6]In this paper I use plural pronouns to refer to a singular generic person of unimportant gender.

Mark Waser (2011) believes treating AIs like criminals will only encourage them to act like criminals; better to assume they'll be benevolent. Eliezer Yudkowsky has written prolifically and thoughtfully on the subject without drawing definite conclusions either way; (Yudkowsky, 2008) is a good overview. I agree that this is a critical issue, especially given the certainty that military funding for AI research will lead to ever-more-autonomous weaponized robots. However, my focus in this paper is on cognitive structures AIs will share with us, not goals they might have that could result in our demise.

3 Irresistible Intuitions

Without further ado, let's try to catalogue some IFIs. In this paper, the focus is on *universal* IFIs, shared by all intelligent agents; and *global* IFIs, accepted by all humans. Global IFIs are marked with a single asterisk; two asterisks are used to indicate a candidate belief with no claim that any population of agents would actually accept it. I will of course provide an argument for each claim of universality.

3.1 Basic IFIs

The first IFI is the paradigmatic example of the genre:

> *My future decisions are free, in the sense that they are uncaused, simply* outside *the world's network of causality, as are, to a lesser degree, objects and people influenced by my decisions.* (1)

Some possible futures are better than (preferable to, more satisfactory than) others. The criteria by which preference is measured include: C_1, C_2, C_3, ... (2)

In the statement of IFI 2, I have schematically included some of the actual criteria, using the labels C_1, C_2, etc. This is to indicate that each agent's criteria are part of its IFI (although there is no implication that an agent can articulate them; they might be embodied in neural nets). It will become clear shortly why IFI 2 is placed here, immediately after IFI 1.

Discussion of IFI 1 gets us into deep into the issue of free will. It is not necessary to disbelieve$_E$ in free will to accept my main point, which is that whether or not you believe$_E$ in it, you can't help but believe$_I$ IFI 1.

This disconnect is classic. It is well put by Galen Strawson (1986, p. 112), who does not believe$_E$ that freedom is possible in a deterministic universe:

> We are, in the most ordinary situations of choice, unable not to think that we will be truly or absolutely responsible for our choice, whatever we choose. Our natural thought may be expressed as follows: even if my character is indeed just something given (a product of heredity or environment or whatever), I am still able to choose (and hence act) completely freely and truly responsibly, given how I now am and what I now know; this is so whatever else is the case—determinism or no determinism.

My contribution is to locate the belief$_I$ in a particular place: the decision models used by intelligent agents. For such models as used by people, see (Johnson-Laird, 1983; Wells and Gavanski, 1989; Goldman, 2006); for their use in AI, see (Forbus,

1984; Ghallab et al., 2004). Knowledge of how decision and simulation take place in the brain is in an embryonic state, but it is already clear that several regions are involved (Baars and Gage, 2010). Decision and action are associated with the dorso-lateral prefrontal cortex (Unterrainer and Owen, 2006). Reasoning about the mental states of other humans (exercising the "theory of mind") involves the temporo-parietal junctions and the posterior cingulate (Saxe and Powell, 2006). Less is known about the neuroscience of simulation of physical systems, but there are preliminary data about perception of causal interactions among objects (Fugelsang et al., 2004).

An impartial spectator could put every entity in the model on the same footing: things are going to be caused to happen or not; or perhaps there are several possible futures with significant probabilities. But when an agent's own behavior is an important causal factor, then its effectors and the things they affect must be placed in a different class from other entities.

For example, suppose an intelligent robot is on the verge of being trapped in a place it wants to escape from. There are two open doors, but they close automatically and the closing mechanism has been set in motion. The robot can reason about the speed and acceleration of the doors purely causally, but it cannot reason about its own body in the same way because it has to choose, among other things, which door to head for. Before it decides to go for one door or the other, the branch point to two futures that might flow from this decision is flagged as "immune from causal analysis."

In decision theory (Raiffa, 1968), situations are represented as tree structures in which nodes occur in layers; *decision* nodes alternate with *simulation* nodes. At a simulation node, the decision maker reasons about the things the world might do

using what simulation tools are appropriate. Sometimes this requires thinking about random possibilities, as in the top-left simulation node of figure 1; sometimes the world may be treated as (locally, approximately) deterministic, as in the top-right simulation node of figure 1, where only one future is considered. At decision nodes, the decision maker reasons about what it *can and should* do. The tree bottoms out in *value nodes,* which summarize the net value of the branches they end. The value of a simulation node is the *expected* value of its children given the probability distribution that governs how it splits. The value of a decision node is the *maximum* value of its children, because the decision maker is "free to choose" and will certainly pick the highest-value option.[7] I do not mean to imply that all intelligent agents must adhere to the tenets of standard decision theory, because humans certainly do not (Kahneman, 2011); but they must draw the distinction between choice and simulation.

Imagine a decision maker that failed to make this distinction, between aspects of the situation under its control and aspects that should be simulated and sampled. Suppose it treated a decision point as if it were just another process to be predicted. Three possible things could happen:

1. Straightforward causal analysis works well. Suppose the agent is a robot sinking to the ocean floor, bound, gagged, and tied to a cinder block. One can predict, without taking any decisions the agent makes into account, how long it will take for salt water, pressure, and battery or air deprivation to destroy it.

2. The agent is in the middle of a deliberation that will actually make a difference

[7]For simplicity, I have left adversarial reasoning and many other complications out of the scenario.

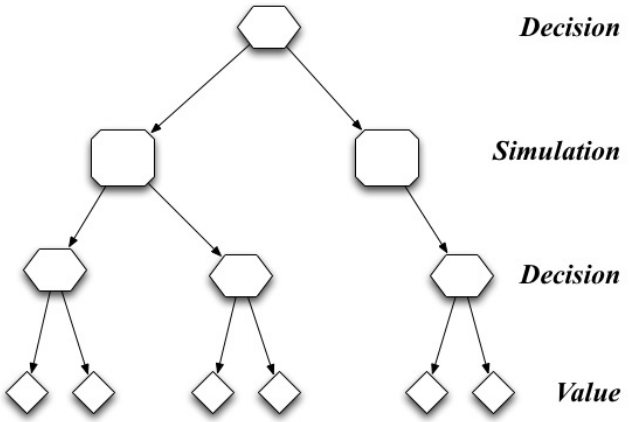

Figure 1: Decision-theoretic scenario

to what it does, but the causal analyzer fails to realize that. Hence its analysis is likely to be wrong.

3. The causal analyzer tries to take the agent's decisions into account. It will get to the point where it has to analyze the analysis in progress and decide how it will come out. This analysis must either get into an "infinite recursion" or revert to case 2.

Cases 2 and 3 require some elaboration.

In case 2, note that even assuming the causal analyzer succeeded in predicting what the agent would do, nothing is actually causing the agent to do that thing. A prediction in a situation where a *choice* is needed is a category error. Predictions do not all by themselves lead to action, even if they are predictions of what one will do. I sometimes have the experience, in a conference room where people are introducing themselves one by one and my mind is wandering, of predicting that

when it's my turn I will say my name; and then realizing that in addition to making the prediction, I actually have to say my name. The prediction that I'm going to speak was based on a simple inductive inference: *Everyone pointed to so far has said their name, so when I'm pointed to I'm going to say mine.* This is a case-2 mistake; the prediction ignores the fact that I am a decision maker. One might envision a decison maker that always takes the predicted action, if any, and carries it out. But this is not a particularly good policy, and in any case what's needed *is* a policy, not a prediction. Mechanical adherence to the policy of following such predictions might be very foolish if, say, I am an undercover spy who doesn't want his name revealed.

What goes wrong when the causal analyzer *does* try to take into account the fact that a deliberation is going on (case 3 above) is the case that is trickiest to grasp. The problem is not that the world and the intelligent agent are too complicated to reason about; nor is there is some special obstacle stemming from a complex agent's being inherently too complex to be understood by itself. Think of it this way: Label the deliberation in progress with the letter D. The deliberator enters a causal-analysis stage D_A which involves trying to analyze D.[8] D's further progress depends on the results of D_A (or it's not necessary and we're back in case 1). But then D_A can discover at most that D will wait for D_A — the very analysis that is happening *now*. It must wait for itself. This is essentially Popper's point in (Popper, 1950a,b).[9]

[8]To remove one possible confusion: D and D_A are event tokens, not event types: particular occasions of the running of processes for making deliberations and predictions.

[9]Alvin Goldman (1968) is able to produce some examples in which an agent's discovery of a prediction of its behavior does not disturb its decision to do exactly what is predicted. But the predictions are produced by someone else using unspecified methods. In each case there is a

To avoid all these problems an intelligent deliberator *must* classify decisions as different from simulations, which is equivalent to modeling the parts of the future that depend on its deliberations as being exempt from causal analysis. This is why everyone has the erroneous belief$_I$ that their future decisions are free, a belief that is the obvious candidate to head the list of irresistible framework intuitions.

Now for IFI 2, another integral component of the agent structure discussed above. The whole point of generating alternative versions of the future is to choose the one that is, all things considered, the best. The structure works only if the agent does normally prefer some futures to others. The belief$_I$ that different futures *have* different values, which I labeled IFI 2, is "wired in" as deeply as the belief$_I$ in my actions being exempt from causal laws.

The way these beliefs$_I$ are embodied is closely analogous to the way the belief that it should get its queen out early is embedded in Dennett's chess program (see section 1). The belief$_I$ that the agent actually makes decisions about futures it actually cares$_I$ about is due to basic design features of its causal-analysis mechanism. In fact, it is a basic feature of the causal-analysis machinery of *any* intelligent agent. Get rid of the embedded belief, and the agent ceases to be intelligent (or perhaps an agent at all).

In (Strawson, 1986, Appendix A), Galen Strawson asks whether a chess program (whose decisions fit the structure illustrated by figure 1 pretty well) could be said to be free. He rejects the idea, unless the chess program was conscious, but asks (p. 319), "What *exactly* is the difference that experience has made? ... What

deliberation in addition to the prediction, whose outcome just happens to coincide with what is predicted. One can conceive of a deliberator that searched for "fixed points" in decision space of this sort, but it's hard to see how it could be very intelligent.

freedom-relevant capacity has experience added?" The question reflects an understandable conflation of two separate capacities an agent might have: the capacity to model itself as an experiencer, and the capacity to model itself as a decider. Both are the sort of self-fulfilling belief that consciousness appears to consist in, but it's only the latter that would turn a chess program into a system with "free" will. (Modelling oneself as an experiencer is the subject of section 3.2.) All it would take to give a chess program as much free will as we have is to give it the capacity to reason about itself *qua* decision maker in addition to reasoning about chess. Note that moral responsibility, or even the idea that an agent (self or other) "could have done otherwise" does not enter into IFI 1; people care about moral responsibility, but other intelligent agents might not.

An artificial intelligent agent, or an intelligent alien from another planet, might and probably will have very different C_i from humans, whose evaluation criteria are probably pretty similar. In particular, I take it as obvious that all people subscribe to this IFI:

> **My own welfare (prolonged, relatively pain-free existence) is one of the principal factors in weighing alternative futures.* (3)

But it is *not* universal, as I will discuss in section 3.3.

I turn now to beliefs about the workings of one's own perceptual systems.

3.2 Perceptual IFIs

The IFIs I discuss in this section are concerned with introspection about the senses. It's useful to begin with some observations about how sensors normally function.

I own an alarm clock that responds to a catalogue of voice commands. You must

first activate it by saying "Hello, Ivee."[10] A remarkable range of other sounds, for example, the opening of a zippered garment bag, may activate it as well, so you have to keep deactivating it by saying, "Cancel." And you often have the irrational urge to ask it, "Just how do the words 'Hello, Ivee" and a zipper sound similar to you?" Irrational not just because the clock's repertoire includes no phrases about sounds and seemings, but because *no signal that ever passes through it,* or, if you prefer, *no interaction with the world that it is capable of,* relates to the issue of the similarity of these two sounds. (Although the question is not vacuous, because if you posed it to the people who built and and programmed the clock, they could eventually provide an answer.)

The next example is closer to home (if overfamiliar to philosophers), and that's the phenomenon of *blindsight,* in which patients with cortical damage have no experience of sight but still can guess accurately where objects are (Weiskrantz et al., 1974; Weiskrantz, 1997).

In both these cases, there is information flow but no ability to introspect about it. We think of such cases as abnormal, but in fact it is the other way around: To be able to introspect about perception, some special "circuitry" (or neural anatomy, or software structure) must exist (Lycan, 1987, 1997). For instance, consider our experience of *stereo disparity,* which arises because of the slight differences in view between the left eye and the right eye. There is a shift in the retinal images of a surface feature as between the two eyes, more of a shift for closer features, giving a clue to depth. Features from the two eyes are matched and their disparities calculated at an early stage of visual processing, as shown by our ablity to see depth in random-dot stereograms (Julesz, 1971), depth which is of course not really there.

[10]Ivee is a trademark of the Ivee Personal Assistant Company.

If an experimenter were to ask you to introspect about how you were misled into seeing depth in such a stereogram you would get nowhere. People have no more introspective access to the feature-matching process than my clock radio has to its sound-matching process.

Contrast those examples of inability to introspect with the phenomenon of apparent size, as described by (Peacocke, 1983, p. 12):

> Suppose you are standing on a road which stretches from you in a straight line to the horizon. There are two trees at the roadside, one a hundred yards from you, the other two hundred. Your experience represents these objects as being of the same physical height and other dimensions; that is, taking your experience at face value you would judge that the trees are roughly the same physical size .... Yet there is also some sense in which the nearer tree occupies more of your visual field than the more distant tree. This is much a feature of your experience itself as its representing the tree as being the same height.

In other words, for whatever reason, "sensory-field height" of a tree is accessible to you in a way that stereo disparity and most other aspects of perception are not.[11]

Having made these distinctions, let me propose the existence of the following IFIs, which hold for those aspects of sensing that an agent *does* have introspective access to. Remember that these are not necessarily true, just believed$_I$.

My perceptions are mediated by sensations. (4)

[11]I thank Aaron Sloman for pointing out to me that our ability to carry out Peacocke-style exercises is surprising and contingent. I am skeptical that that a theory such as that of O'Regan and Noë 2001, in which introspection is inherent in behavior, can be made to work.

Sensations have spatiotemporal structure, but also nonrelational, qualitative aspects that I am acquainted with directly. (5)

I make distinctions among perceived things and situations by making distinctions among the sensations they cause, using their spatial structures and the qualia *of those sensations to know the differences among them.* (6)

Obviously, there is no fact of the matter about whether a list like this is numerically correct. Where I write three IFIs, someone else might write two or four. But the belief$_I$ that between reality and me there lies some sort of appearance, the deliverance of the senses, seems irresistible (IFI 4).

The traditional sense systems are distinguished by their association with *phenomenal fields,* the visual field being the paradigmatic example. When I introspect about sight, there seems to be a two-dimensional film of experience between me and the world, through which all visual information about the world flows. Similarly, there is a three-dimensional field of sound sensations that mediate my hearing of events and objects near me. The phenomenal fields are aligned with the sensory systems that underlie them. The alignment is not perfect; the taste field is constructed using information from the taste sensors in the tongue and smell sensors in the nasal cavity.

Whether or not natural selection cares, it is a fact that the possession of a phenomenal field for a sense gives its possessor learning opportunities it could otherwise not have. Having a visual field provides an agent with a model$_I$ of error in visual sensing, namely, that something *appeared* to be one way, when *in fact* it was otherwise. Similar remarks could be made about the other four senses, but not, for instance, about pain. Although there are physiological facts about whether pain

reports apparently originating at a body part are veridical (or actually originated at some distal point), we never *model* ourselves as seeming to have a sore toe but not really having one.

This is what makes IFIs 4 through 6 universal: any intelligent agent with episodic memories of sensory events will organize them into *appearance* vs. *reality*, and the appearance side will be a phenomenal field. That is, it will have an optional spatiotemporal aspect, organized as a changing two- or three-dimensional array, and a qualitative aspect, organized using however many dimensions are appropriate to that sense. Of course, for the human race, it requires ingenious psychophysical experiments to infer the dimensionality of a quality space; people have no introspective access to this information. An artificial agent might not have this limitation. But it would still have the distinction between the spatiotemporal organization of the field and the qualitative aspect.

It may seem to be a contingent fact about humans that our phenomenal fields have spatiotemporal qualities and those other *qualia* that are usually described as nonrelational, private, and ineffable, but a moment's contemplation will show that those are the only choices: the units that are used in stating physical laws, and the other, secondary qualities. Imagine a robot with binary vision: all its eyes give it is an array of ones and zeroes, a one corresponding to an amount of light greater than some threshold. If this robot can introspect about its visual experience, it might have access to a two-dimensional field; and the smallest entity it could be aware of would be a single pixel, similar to some, and different from other, pixels around it.[12] If you asked it, "How do you know that this pixel differs from that one?," its answer

[12]I do not mean to predict that intelligent robots, if such there ever are, will be able to introspect about single pixels. They probably would be as bad as we are at a task like that.

would have to be that it has a different quality. In other words, it can make use of publicly available dimensions, the primary qualities, to talk about the layout of the visual field and the objects in it, but it would have to invoke alternative dimensions, or secondary qualities, to talk about the distinctions between the two sort of pixel.

The aural field serves the purpose of locating objects in space, but in humans has come to have aesthetic properties as well. I believe it would not be too difficult to build a device to detect the sound of trucks backing up, perhaps to aid the hearing-impaired. Similarly, one could (with further technological progress) help klezmer haters stay a safe distance from that sound with a device for detecting the presence, direction, and range of any klezmer bands in the vicinity. But obviously what we normally do when listening to a band is attend to the sensations rather than perform target acquisition on their source. When and why this practice evolved is unknown. We can do it with any sound (for instance, we might come to believe that truck beeps are needlessly loud and salient), but music is sound produced for the specific purpose of attending to the aural field it creates.

There is no reason why the spatial dimensions of the phenomenal field must correspond to those of real space. Consider a creature for which color is more important than shape; it might have a phenomenal field whose spatial dimensions correspond to its species' color space (whatever that might be), and whose points were occupied by qualia corresponding to the direction where the brightest sample of that color is to be found. Its introspections might reveal that "x, y, z in color space feels leftish." An example closer to home is our visual field, whose two dimensions do not, after all, correspond to any manifold in physical space (Noë, 2002).

3.3 Parfitean IFIs

Some of the thought experiments devised by philosophers seem to be aimed at setting us free from some seemingly irresistible, but groundless, intuitions. Derek Parfit's classic work *Reasons and Persons* (1984) provides several examples. To take one, he argues in chapter 11 that personal identity over time is simply a matter of resemblance. My future self is whoever has my memories, under the reasonable, and normal, assumption that that there is a continuous series of selves joining me now with me in the future. His argument for this conclusion is fairly convincing. But what I want to focus on is his claim that by focusing hard on the scenario he can, for instance, imagine deciding to step into a teleportation machine that will create a complete description of his brain and body, *destroying them in the process,* then transmit the description to, say, Mars, where it will be used to reconstitute them.[13] Thus he claims to refute

> **The physical annihilation of "me" counts very heavily (weight W) against any projected future in which it occurs.* (7)

Parfit admits that he has to work himself into a special mood to avoid this belief; it keeps coming back (Parfit, 1984, p. 279–280). In other words, it's an IFI, a corollary of IFI 3. I will call the IFIs picked out by this method *Parfitean.*[14] (In IFI 7 I've once again schematized the actual weight, which varies from agent to agent. In people who are severely depressed or in constant pain, W can become negative, but this does not make the claim vacuous, because $W > 0$, and $\gg$ the weights for most other criteria, for the great majority of people.[15])

[13] The teleportation scenario was first used by (Williams, 1970).

[14] Although each has the support of many other philosophers as well.

[15] Although Parfit's thought experiment does suggest another odd possibility: Would a depressed

IFI 7 is shared by almost all humans, but it is not universal, because it is easy to imagine a robot that doesn't fear its own destruction under the right circumstances (a *truly* smart bomb, for instance).

Parfit devotes chapter 14 of (Parfit, 1984) to showing that it is self-defeating or irrational to make decisions in order to maximize your total happiness over your lifespan. The argument hinges on a corollary of his conclusions about personal identity cited above, namely, that you, *qua* you, *don't actually exist* for any longish period of time, because you change so much over time that your lifespan actually consists of a succession of different persons.[16] Therefore, in considering the goals of your "self" sufficiently far in the future, there is no reason to attach more weight to them than to those of some other agent in a similar position. The argument is completely convincing in a sense; the only problem with it is really a problem with (Parfit, 1984) in general: he assumes that an argument$_E$ is all you need. He is trying to work from within the purely scientific point of view, but from that point of view the self doesn't exist at all; all that is observable are agents that *believe*$_I$ they are selves. (Nagel (1986) calls this "the view from nowhere." The world of IFIs is profoundly different from the scientific world; unfortunately, as I briefly discuss in section 4, we must attempt to live with one foot in each world.

The self concept includes the feature that the self we believe$_I$ in will exist for (at least) as long as there exists a body that it controls. This claim is actually subject to proof. First, realize that when I say "body," you needn't visualize a compact, person-sized automaton. The body could be as simple as an I/O stream connecting the agent to a chat room, or as dispersed as an automobile-traffic control system.

person believe their suffering would be ended by stepping into a teleportation machine, knowing that a person identical to them would resume suffering at the other end?

[16]Perhaps Parfit's discovery should probably be credited to the Buddha. See also (Nagel, 1978).

But if an agent A's projections of the future do not mark parts of a projection as corresponding to A, then there is no sense in which A has a model of itself at all. If it has no model of its future, then it has no model of itself *as* a self. If we take the contrapositive of this conclusion, we infer that *any conscious agent will believe in the persistence of itself through time.* Or, stated from within the agent's belief$_I$ system:

I will continue to exist. (8)

It is impossible to conceal from a truly intelligent agent that 8 cannot hold true over an infinite future. This realization has been cited as the beginning of philosophy, suggesting that artificial intelligent agents might join in the philosophical discourse.

It might seem like a parlor trick to take a conclusion of the form "Any conscious agent must *believe*$_I$ P" and transmogrify it to "P (from within the agent's belief system)" and thence to, simply, "P." But if we, that is, I, that is, *you* are the agent in question, you can't stay out of the zone where P is a live belief$_I$ much longer than you can hold your breath.

One final Parfitean IFI:

It is preferable for a painful episode to be in one's past rather than one's future. (9)

This IFI is not of central importance, and it is easy to misread it, as I will discuss shortly. But Parfit condemns it in favor of the "Timeless" point of view (p. 174) that treats as equivalent past and future pleasure and pain. He grants the role of evolution in wiring opposing views into us, but fails to appreciate the importance of this observation. In fact, I argue that IFI 9 would be wired into any intelligent

agent.[17] It is a simple consequence of the asymmetry of future projection and memories of the past. The whole point of future projection is to weigh alternative versions of the future, and part of the definition of pain is that a large negative weight is attached to a prediction of pain when doing the comparison. The purpose of episodic memory is not so clear, but obviously there is nothing like a weight in the same sense attached to a memory. There are other sorts of values we attach to memories — which episodes we recall with nostalgia or guilt, or value for one purpose or another. A painful episode might end up with a positive weight, possibly *because* it involved pain. Whatever the point of these systems of valuation, they are obviously at best only loosely coupled to the system for deciding which alternative futures to strive for.

Parfit is aware of this argument, but interprets it as a pleading for a point of view rather than a specification for our "wiring diagram." So he argues that it would be possible, even rational, for a person to adopt a "Timeless" attitude, in which past pleasures and pain are added in to one's balance sheet the same way future ones are, all weighted the same, regardless of the time at which they occur (Parfit, 1984, p. 174). But imagine an agent that really weighed timelines that way. It would not take long before the contribution from the past, being summed over a growing trail

[17] An exception must be made for agents that can't feel pain. Today's robots don't experience pain, or anything else, but by the time intelligent robots exist, if they ever do, sensor technology will surely have reached the point where it will make sense to equip the machine with damage sensors throughout its body, and it will conceptualize signals from those sensors as pain (McDermott, 2001). But one might set up an intelligent city-management system, and it might not be so important for it to react to the destruction of a traffic light with the level of urgency we attach to bodily injury. It's hard to say what determines whether an injury should be painful; why do animals have no pain sensors in their brains?

of memory, completely swamped the contribution from the future projection, which would always be the same short interval that we manage to foresee. The agent would become more and more indifferent about its decisions as the change in its overall objective function came to depend less and less on them. It would quickly cease to be an *agent* at all, instead becoming a passive observer.

Let me forestall a possible misreading of IFI 9. It might seem to follow that

> ***Given an unavoidable but procrastinatable painful episode in one's future, it is better not to procrastinate.* (10)

(Recall that two asterisks mark propositions that I am not claiming to be anybody's IFI.)

The "proof" of 10 would be that, since, all other things being equal, it is better to have a painful episode in one's past, the future in which the episode is in one's past quicker will have a longer span of time with the pain safely in the realm of memory rather than anticipation. The reason this reasoning is fallacious is that it neglects discounting of future pleasures and pains. See (Ainslie, 2001) for a thorough explanation of the alternative discounting functions and their ramifications. So there are people who find 10 compelling, but by no means all.

4 Conclusions

All intelligent agents, including those the human race may soon create, are or will be subject to beliefs that are difficult or impossible to question, without regard to whether they are actually true. These are not isolable beliefs that it is possible to tiptoe around, but wired-in operating assumptions, for which I use the notation "beliefs$_I$." Hence they are *irresistible framework intuitions*, abbreviated *IFIs*. While

all intelligent agents will have some IFIs, they won't be the same as the ones humans live with. I use the term *universal IFIs* for intuitions that all intelligent agents would have to have, and I propose some. Here are a few:

- My future decisions are exempt from causal laws. (IFI 1)
- My perceptions are mediated by sensations. (IFI 4)
- It is preferable for a painful episode to be in one's past rather than one's future. (IFI 9)

One might wish that the list included some moral intuitions, but it is just too easy to imagine amoral intelligent robots.[18] Indeed, the military seems to be on track to building amoral autonomous robots with the capacity to kill people (Arkin, 2009; Singer, 2009); how intelligent they will get is not clear.

Although these observations are as of third parties, a race under observation, what makes them interesting is that they are about *us,* that is, about you and me. If one of our IFIs is actually false, then we are in the position of finding their negations obviously true and at the same moment literally unbelievable. We live with antinomies such as:

- Everything is subject to causal laws; *vs.*: My future decisions are exempt from causal laws (IFI 1).
- Personal identity is just a matter of memory and similarity through time; *vs.*: I will continue to exist (IFI 8).

[18]It is seemingly easy to imagine evil robots, but this question gets into issues about the ability of robots to make genuine moral decisions, where there is room for skepticism (McDermott, 2011b).

- My suffering pain at time t_1 is no better or worse than my suffering pain at time t_2, if t_1 and t_2 are only days apart; *vs.*: Pain tomorrow is much worse than pain yesterday (IFI 9).

As P.F. Strawson put it, there are "two different standpoints from which human behavior may be viewed: for short, the 'participant' versus the 'objective,' the 'involved' versus the 'detached.' ... One cannot be wholeheartedly committed to both at once" (Strawson, 1985, p. 36).

Note that it is only since the beginning of the scientific age that mankind has had to suffer from these continual clashes. Without the Enlightenment, most of us would happily live on the second horn of each dilemma without noticing that the dilemma existed. How can we accept such a seemingly intolerable conclusion, that ever since science began churning out indubitable truths that deny cherished IFIs, our belief systems have become inherently contradictory and unstable? From a contradiction it is possible to infer anything, so one might expect people to be unable to function at all. Instead we seem to alternate between belief systems, using the one that is most appropriate to the situation.

Even working scientists live outside the scientific framework most of the time. Physicists make free decisions about how to conduct experiments based on physical theories that leave no room for that kind of freedom, as do social scientists studying the "illusion of conscious will" (Wegner, 2002). Cognitive scientists seek to explain phenomenal consciousness as a useful illusion while listening to music on their iPods. It is hard to see these people as compartmentalizing their lives in, say, the way a district attorney might be imagined to do who goes to church on Sunday to pray for the forgiveness of sinners. Instead, the scientist oscillates, sometimes minute-by-minute, from one side of the IFI membrane to the other. We can't escape our

beliefs$_I$ for very long, even those we sincerely believe$_E$ to be false.

Acknowledgements: Thanks to David Gelernter, Karsten Harries, and especially the anonymous referees for comments on earlier drafts of this paper. I received valuable suggestions from Joshua Knobe and Brent Strickland. Portions of this paper were presented in a Fall 2010 talk before the Yale Cognitive Science program, whose members I thank for feedback, especially Tamar Szabo-Gendler. Unfortunately, for the many flaws remaining in the paper there is no one to blame but the author.

References

Abbott, L. and Sejnowski, T. J., editors (1999). *Neural Codes and Distributed Representations: Foundations of Neural Computation*. MIT Press, Cambridge.

Ainslie, G. (2001). *Breakdown of Will*. Cambridge University Press.

Akins, K. (1996). Of sensory systems and the "aboutness" of mental states. *J. of Phil*, 93(7):337–372.

Arkin, R. C. (2009). *Governing Lethal Behavior in Autonomous Robots*. Chapman and Hall/CRC, Boca Raton.

Baars, B. and Gage, N. (2010). *Cognition, Brain, and Consciousness: Introduction to Cognitive Neuroscience (2nd edition)*. Academic Press, Amsterdam.

Baars, B. J. (1988). *A Cognitive Theory of Consciousness*. Guilford Press, New York.

Baars, B. J. (1997). *In the Theater of Consciousness: The Workspace of the Mind*. Oxford University Press, New York.

Baum, E., Hutter, M., and Kitzelmann, E., editors (2010). *Proc. Conf. on Artificial General Intelligence*. Atlantis Press, Amsterdam.

Belfiore, M. (2007). Carnegie takes first in DARPAs urban challenge. *Wired*.

Chalmers, D. (2010). The singularity: A philosophical analysis. *Journal of Consciousness Studies*, 17(9–10):7–65.

Dehaene, S. and Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, 79:1–37.

Dennett, D. C. (1977). A cure for the common code? *Mind*. (Reprinted in (Dennett 1978c), pp. 90–108.).

Dennett, D. C. (1984). *Elbow room: The Varieties of Free Will Worth Wanting*. MIT Press, Cambridge, Mass.

Dennett, D. C. (1991). *Consciousness Explained*. Little, Brown and Company, Boston.

Dennett, D. C. (2003). *Freedom Evolves*. Viking Books.

Denyer, N. (1981). *Time, Action, and Necessity*. Duckworth, London.

Falkenhainer, B., Forbus, K. D., and Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, 41(1):1 – 63.

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., Lally, A., Murdock, J. W., Nyberg, E., Prager, J., Schlaefer, N., and Welty, C.

(2010). Building Watson: An overview of the DeepQA project. *AI Magazine*, 31(3):59–79.

Fodor, J. (1975). *The Language of Thought*. Thomas Y. Crowell, New York.

Forbus, K. (1984). Qualitative process theory. *Artificial Intelligence*, 24:85–168.

Forbus, K., Ferguson, R. W., and Gentner, D. (1994). Incremental structure-mapping. In *Proc. Conf. of the Cognitive Science Society*, volume 14, pages 313–318.

Fugelsang, J. A., Roser, M. E., Corballis, P. M., Gazzaniga, M. S., and Dunbar, K. N. (2004). Brain mechanisms underlying perceptual causality. *Cognitive Brain Res*, 24:41–47.

Gallistel, C. R. (1990). *The Organization of Learning*. MIT Press, Cambridge, Mass.

Gendler, T. S. (2008). Alief and belief. *Journal of Philosophy*, 105(10):634–663.

Gendler, T. S. (2009). Alief in action (and reaction). *Mind & Language*, 23(5):552–585.

Ghallab, M., Nau, D., and Traverso, P. (2004). *Automated Planning: Theory and Practice*. Morgan Kaufmann Publishers, San Francisco.

Goertzel, B., Hitzler, P., and Hutter, M., editors (2009). *Proc. Conf. on Artificial General Intelligence*. Atlantis Press, Amsterdam.

Goldman, A. (1968). Actions, predictions, and books of life. *American Philosophical Quarterly*, 5(3):135–151.

Goldman, A. I. (2006). *Simulating Minds:The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford University Press.

Harman, G. (1973). *Thought*. Princeton University Press.

Harman, G. (2000). *Explaining Value: and Other Essays in Moral Philosophy*. Clarendon Press, Oxford.

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*. MIT Press, Cambridge, Mass.

Hofstadter, D. R., editor (1995). *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. Basic Books, New York. (by Douglas Hofstadter and the Fluid Analogies Research Group).

Inwagen, P. V. (1983). *An Essay on Free Will*. Oxford University Press, New York.

Johnson-Laird, P. N. (1983). *Mental Models*. Harvard University Press, Cambridge, MA.

Julesz, B. (1971). *Foundations of Cyclopean Perception*. University of Chicago Press.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.

Kanwisher, N., McDermott, J., and Chun, M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci*, 17(11):4302–11.

Kurzweil, R. (2005). *The Singularity is Near*. Viking, New York.

Lakoff, G. and Johnson, M. (1980). *Metaphors we Live By.* Chicago .University Press.

Lycan, W. G. (1987). *Consciousness.* MIT Press, Cambridge, Mass.

Lycan, W. G. (1997). Consciousness as internal monitoring. In Block, N., Flanagan, O., and Güzeldere, G., editors, *The Nature of Consciousness: Philosophical Debates*, pages 755–771. MIT Press, Cambridge, Mass.

Mackie, J. (1977). *Ethics: Inventing Right and Wrong.* Penguin Books, Harmondsworth, Middlesex.

McCormick, I. A., Walkey, F. H., and Green, D. E. (1986). Comparative perceptions of driver ability— a confirmation and expansion. *Accident Analysis and Prevention*, 18(3):205–208.

McDermott, D. (2001). *Mind and Mechanism.* MIT Press, Cambridge, Mass.

McDermott, D. (2011a). A little static for the dynamicists. *Int. J. of Machine Consciousness*, 3(2). (Review of *Embodiment and the Inner Life,* by Murray Shanahan.).

McDermott, D. (2011b). What matters to a machine? In Anderson, S. and Anderson, M., editors, *Machine Ethics*, pages 88–114. 2011 Cambridge University Press.

McKay, R. T. and Dennett, D. (2009). The evolution of misbelief. *Behavioral and Brain Sciences*, 32(6):493–561.

Metzinger, T. (2003). *Being No One: The Self-Model Theory of Subjectivity.* The MIT Press, Cambridge, MA.

Nagel, T. (1978). *The Possibility of Altruism*. Princeton University Press, Princeton. (Originally published 1970 by Oxford University Press).

Nagel, T. (1986). *The View from Nowhere*. Oxford University Press.

Newell, A. (1969). Heuristic programming: Ill-structured problems. In Aronofsky, J., editor, *Progress in Operations Research*, pages 361–414. Wiley, New York.

Nietzsche, F. (1886). *Beyond Good and Evil*. The Modern Library, New York. In Walter Kaufmann (ed.) 1968 *Basic Writings of Nietzsche*. (Translated by Walter Kaufmann).

Nisbett, R. E. and Wilson, T. D. (1977). Telling more than we can know: Verbal reports of mental processes. *Psychological Review*, 84(3):231–259.

Noë, A. (2002). Is the visual world a grand illusion? *J. of Consciousness Studies*, 9(5-6):1–12.

Omohundro, S. M. (2008). The basic AI drives. In *Proc. First Conf. on Artificial General Intelligence*, pages 483–492. Amsterdam: IOS Press.

O'Regan, J., Noë, A., et al. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24(5):939–972.

Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.

Peacocke, C. (1983). *Sense and Content: Experience, Thought, and their Relations*. Clarendon Press, Oxford.

Popper, K. R. (1950a). Indeterminism in quantum physics and in classical physics. *British J. for the Phil. of Science*, 1(2):117–33. part I.

Popper, K. R. (1950b). Indeterminism in quantum physics and in classical physics, part ii. *British J. for the Phil. of Science*, 1(3):173–95.

Raiffa, H. (1968). *Decision Analysis: Introductory Lectures on Choices under Uncertainty*. Addison-Wesley.

Russell, S. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach (3rd edition)*. Prentice Hall.

Ryle, G. (1949). *The Concept of Mind*. University of Chicago Press.

Saxe, R. and Powell, L. J. (2006). Itś the thought that counts: specific brain regions for one component of theory of mind. *Psychological Science*, 17(8):692–699.

Schiffer, S. (1972). *Meaning*. Oxford University Press.

Schmidhuber, J., Thorisson, K. R., and Looks, M., editors (2011). *Proc. Conf. on Artificial General Intelligence*. Spring Verlag. (Lectura Notes in AI 6830), Berlin.

Schwitzgebel, E. (2008). The unreliability of naive introspection. *Phil. Rev*, 177(2):245–273.

Shanahan, M. (2010). *Embodiment and the Inner Life*. Oxford University Press.

Sidgwick, H. (1907). *The Methods of Ethics*. Macmillan and Company, London. (seventh edition). (Reprinted 1962 by the University of Chicago Press.).

Singer, P. (2009). *Wired for War: The Robotics Revolution and Conflict in the 21st Century*. The Penguin Press, New York.

Sloman, A. and Chrisley, R. (2003). Virtual machines and consciousness. *J. Consciousness Studies*, 10(4–5):6–45. Reprinted in (Holland 2003), pp. 133–172.

Smart, J. (1961). Free will, praise, and blame. *Mind*, 279:291–306.

Steels, L. and Brooks, R., editors (1995). *The Artificial Life Route To Artificial Intelligence: Building Embodied, Situated Agents*. Lawrence Erlbaum Publishing, Mahwah, N.J.

Strawson, G. (1986). *Freedom and Belief*. Clarendon Press, Oxford.

Strawson, P. (1962). Freedom and resentment. In *Proc. British Acad*, volume 48, pages 187–211. (reprinted in (Strawson 1974) and (McKenna and Russell 2008), pp. 19–36).

Strawson, P. (1985). *Skepticism and Naturalism: Some Varieties*. Columbia University Press, New York. (The Woodbridge Lectures 1985).

Tinbergen, N. (1951). *The Study of Instinct*. Clarendon Press, Oxford.

Tooby, J. and Cosmides, L. (1992). The psychological foundations of culture. In H. Barkow, L. C. and Tooby, J., editors, *The Adapted Mind*, pages 19–136. Oxford University Press, New York. 1992.

Tooby, J. and Cosmides, L. (2005). Conceptual foundations of evolutionary psychology. In Buss, D., editor, *The Handbook of Evolutionary Psychology*, pages 5–67. Wiley, Hoboken. 2005.

Unterrainer, J. M. and Owen, A. M. (2006). Planning and problem solving: From neuropsychology to functional neuroimaging. *J. of Physiology – Paris*, 99:308–317.

Waser, M. (2011). *Rational universal benevolence: Simpler, safer, and wiser than "Friendly AI."*. *Proc. Conf. on Artificial General Intelligence* **4**.

Wegner, D. M. (2002). *The Illusion of Conscious Will*. MIT Press, Cambridge, Mass.

Weiskrantz, L. (1997). *Consciousness Lost and Found: A Neuropsychological Exploration*. Oxford University Press.

Weiskrantz, L., Warrington, E., Sanders, M., and Marshall, J. (1974). Visual capacity in the hemianopic field following a restricted occipital ablation. *Brain*, 97:709–728.

Wells, G. L. and Gavanski, I. (1989). Mental simulation of causality. *J. of Personality and Social Psych*, 56(2):161–169.

Williams, B. (1970). The self and the future. *Philosophical Review*, 79(2):161–180. (Also in Williams 1976, pp. 46–63.).

Yudkowsky, E. (2008). Artificial intelligence as a positive and negative factor in global risk. In Bostrom, N. and Cirkovic, M., editors, *Global Catastrophic Risks*, pages 308–343.