# Social intelligence and interaction

## Expressions and implications of the social bias in human intelligence

EDITED BY
ESTHER N. GOODY
*University of Cambridge*

CAMBRIDGE
UNIVERSITY PRESS
1995

STEPHEN C. LEVINSON

# 11 Interactional biases in human thinking

The human mind is something of an embarassment to certain disciplines, notably economics, decision theory and others that have found the model of the rational consumer to be a powerful one. (Schelling 1988: 353.)

## Background

This chapter sets out to weave an improbable web through such topics as animism, common tendencies in the purchase of soap powder, extra-terrestrial lifeforms, the phrase 'the whatdoyoucallit', and the theory of communication. The thread, if it doesn't break, is the theme of a systematic bias in human thinking, in the direction of interactive thinking (E. Goody's *anticipatory interactive planning* or AIP). Because the argument is somewhat indirect, let me state the thesis right here in the beginning in semi-syllogistic form:

1. Communication is *logically* impossible
2. Nevertheless we humans can communicate
3. Therefore, we must use *non-logical* heuristics and a special form of reasoning to bridge the gap
4. For communication to work routinely, these heuristics must be dominant in our thinking all the time
5. Therefore, these heuristics spill over to bias our thinking in non-communication domains.

As in the famous conclusion to Wittgenstein's *Tractatus*, where we are advised not to think that which cannot be thought, so there is a certain paradox in thinking about biases in human thinking. (You can climb outside human thought, Wittgenstein hinted, just so long as you throw the ladder away and climb quickly back in.) We can only do so with real confidence, perhaps, where we can discern an indubitably correct way of thinking, guaranteed by the laws of mathematics or logic, from which human thinking tends to deviate. One such area is human judgement about uncertain events, and it is here that there has grown, largely

through the efforts of Tversky and Kahneman, a rich literature on biases in human thinking. I am a complete novice in this field, but I can't help considering that it might offer rich pickings for those who discern an underlying human preoccupation with social interaction as the evolutionary source of human intelligence. This then is an entirely exploratory foray out of the theory of communication into a neighbouring field. I am not optimistic that it will be well received in that neighbouring field, but interdisciplinary activity has always been a risky enterprise.

## I Interactional intelligence, coordination and communication

### *Interactional intelligence*[1]

In an engaging book (*Frames of Mind*, 1985), Howard Gardner argued forcefully for the diversity of kinds of human intelligence, using a range of evidence from psychological theory, neurology, case studies of cultural, personal hyperachievements, so-called 'idiot savants', and so on. Amongst the specialized, compartmentalized intelligences, he listed linguistic, musical, logico-mathematical, spatial, kinaesthetic and personal intelligences. Within the latter, he includes what I would choose to isolate as *interactional intelligence*, and he lists as evidence for such a specialized skill the special role of the frontal lobes of the human brain. Persons with frontal lobe damage of various kinds exhibit different but related inabilities: 'No longer does the individual express his earlier sense of purpose, motivation, goals, and desire for contact with others; the individual's reaction to others has been profoundly altered, and his own sense of self seems to have been suspended' (1985: 262). Conversely, patients with massive brain damage to other areas who retain fully functional frontal lobes – like Luria's 'man with the shattered world' – retain the capacity to plan actions and to relate to others to a surprising degree. Similarly, one can point to autistic (and perhaps schizophrenic) patients, who often show signs of unimpaired reasoning ability, but who cannot relate to others; and conversely to reasoning-impaired individuals (like those with Down's syndrome or Alzheimer's) who seem to retain great interpersonal skills.

Making due allowance for the lay misuse of neurological and pathological data, there is here the same kind of range of suggestive evidence for a specific interactional intelligence as there is for other specialized human skills. One should add to this further evidence from the cross-cultural study of interaction: although still in its infancy, and still largely unpublished, such work would seem to establish that there are striking universals in interactional organization, facts compatible with a theory of the biological basis for interactional skills. Studies with infants strongly



suggest such an innate basis under rapid maturation: newborn infants are subtly adaptive to the caretaker's presence and handling, and by two months the child already displays 'a rich repertoire of expressive behaviours . . . combined with ready orientation of the gaze to or away from the mother's face and immediate response to her signs of interest and her talking' (Trevarthen 1979b: 541). That biological basis for interactive skills is further attested to by a wide range of facts about human perception, for example our hearing is acute precisely in the range of wavelengths where speech is broadcast (rather than being specialized like, say, the owl's auditory system, to the noises of prey).[2] Similarly, there is considerable evidence for a specifically human neurological specialization for face recognition, implying the fundamental importance of human face-to-face interaction in human phylogeny.[3] And of course all the physiological, neurological and ontogenetic foundations for language point in the same direction.

The theory of multiple intelligences should not, though, be equated with the modular theory of mind *à la* Fodor (1983); the latter is a particular theory about how specific specialized skills or 'modules' fit together with general thought processes to form a computational whole. The Fodorean requirements for modularity seem altogether too strong to be correct even for linguistic ability taken as a whole (although they plausibly hold for specialist subsystems, like segment recognition), because language understanding necessarily involves general thought processes. In the same way, interactional intelligence (for reasons that will become clear) would have to involve central processing and could not therefore be remotely 'modular' in the Fodorean sense. Nevertheless the skills that jointly make up interactional intelligence seem to be connected intimately enough to make up a package of abilities that can suffer simultaneous neurological impairment.

In this chapter, I shall assume that there is such a form of intellect as an interactional intelligence, and my central concern is whether we can detect a systematic bias in human thinking in other domains which might be attributed to the centrality of interactional intelligence in our intellectual makeup. In order to explore this bias (if such it is), we will need to have some characterization of the central properties of interactional intelligence, which I will attempt to provide.[4]

### *Anecdotal evidence in favour of an interactional bias in human thinking*

Those who would like to replace *Homo ludens*, *Homo hierarchicus* and other such creatures and caricatures with *Homo interactans* (not a possible Latin formation unfortunately,[5] but much more plausible) can find much

anecdotal evidence for such a chap. One of the things that struck Victorian observers (like Fraser and Levy-Bruhl) of 'primitive' peoples was that their world is apparently pervaded by mystic forces in para-human form. Natural causes are mere means subtly utilized by witches, sorcerers, spirits, gods and demons. It is as if the perceptible natural world were a stage set, manipulated by supernatural agents always in interaction with man. Although later ethnographic research (as with Evans-Pritchard's (1937) classic work on Azande witchcraft) has shown us how systems of witchcraft and sorcery have an irrefutable internal reason, make sense in a world imbued with the primacy of social relationships, and so on, it has not thereby made the central problem of such intellectual genera disappear – namely, why we as a species seem predisposed to such intellectual systems, even when they are not socially reinforced or are contrary to our own ideas about real knowledge (as with the astrological systems of early modern astronomers like Tycho Brahe or even Newton[6]). That natural science and magical systems have not only coexisted but often mutually reinforced one another is now a commonplace of the history of science (see, e.g., Lloyd 1979). Scientists often operate (like Watson and Crick) as if nature were a book to be read, a message to be decoded, a syntax to be parsed, a mode of thought that harmlessly enough might be held to presuppose a writer, a coder, a puzzle-setter rigging things behind the perceptible veil.

If scientists are sometimes covert magicians and animists, so of course are children. Piaget (1929) found that children imbue some inanimate objects with intentions, feelings and knowledge, and although later work by Trevarthen and others has shown that very young infants distinguish interactional partners from other kinds of objects for purposes of communication, yet there seems to be a residual blurring of the distinction in the belief world of the child (Gelman and Spelke 1981: 56). One is reminded too of Vygotsky's views about 'inner speech', and indeed the role that an imaginary interlocutor plays in adult thinking.[7]

Other areas where animistic and interactional thinking abound are not hard to find. Consider for example Kahneman and Tversky's finding that experimental subjects treat random processes as if the processes themselves are acting to achieve their own randomness: 'Idioms such as "errors cancel themselves out" reflect the image of an active self-correcting process. Some familiar processes in nature obey such laws ... The laws of chance, in contrast, do not work that way' (Kahneman *et al.* 1982: 24). Economists are often puzzled by the odd purchasing behaviour of consumers – why do they often just buy the most expensive soap-powder? My strategy is to buy the cheapest; my wife's to buy the most expensive. I operate with a vision of some mean, cheating fellow filling different cartons all with the same rotten stuff; she operates with the



vision of the old-fashioned but always reliable and trustworthy owner of the corner-store, whose goods are always more expensive but worth it. The moral is that it's hard to dehumanize even soap-powder. In 1959 astronomical observers started picking up patterned radio signals from outer space. Someone had the idea that extraterrestrial intelligence was trying to contact us: suddenly the signals were being scrutinized in a quite different way, no longer as 'natural signs' of distant physical events, but as 'signals' coded in such a way that any intelligent receiver should be able to decode them.

Presuming an interactor in the inanimate world is one kind of striking conceptual 'error' in human thought; another, less obvious, is the tendency to think of social agglomerations as human actors: we talk happily of what Russia intends in the Baltic, how it will react to NATO, or respond to Islamic fundamentalism. Diplomatic protocol is based on the same principles as interactional politeness (Brown and Levinson 1987); game theory is applied equally to the moves of military or economic conglomerates as to the moves of individual players of parlour games; historians talk in terms of the will of nations. In fact, of course, human agglomerations are propelled through history largely by forces beyond their control: the Russian dismantling of the Iron Curtain may have been no more intentional than an earthquake. Such animistic thinking can have pernicious consequences: we may detect threats where none exists, interpret delayed responses as reluctant or hesitant in character, and find strategic intentions attributed to our non-intentional collective 'actions'.

And so on. There is room enough in the natural history of human belief systems for much speculation about a bias towards the assumption of a world constructed out of human interaction with human and super-human agents. But we seek for some less Victorian level of speculation.

## *Properties of an interactional intelligence*

Human interaction is clearly characterized by an inordinate concern with the implications that an actor's actions have for other actors' expectations, emotions, self-esteem, social status – in short it takes place within a highly structured and often restrictive set of social relationships which permeate the most intricate details of interactional patterning. Nevertheless I want to abstract away from that social matrix, to ask about the underlying *conceptual abilities* that make social interaction possible.[8]

The properties exhibited by human interactants are (from an ethological perspective at least) extraordinary in a number of ways – exactitude of timing, complexity of response, layeredness of meaning, and so on.[9] It is obvious enough that interactional capacity relies on a number of core abilities: the ability to make models of the other, to 'read' the intentions

behind actions, to make rapid interactional moves in an ongoing sequence of actions structured at many levels. But what I think perhaps has not been appreciated is the *computational complexity*, indeed *intractability*, of some of these processes, which is what I want to highlight here.

The computational feat is well illustrated by the ability of humans to defy the laws of chance: to coordinate mutual actions even when unable to communicate with one another.

Schelling, in the *Strategy of Conflict*, reports on some informal experiments that showed that, roughly nine times out of ten, subjects can coordinate plans without communication (1960: 54ff.). Subjects were given a joint goal, but then had to independently work out which means the others would use to solve it, and to choose just that same means. The kinds of problems solved were (a) to think of the same number, the higher the number the higher the reward; or (b) go somewhere determinate in a city to which the other party will also go simply by knowing that each is trying to select the same location and the same time for a meeting (*ibid.*). Subjects coordinated on the number one million, or on the information booth in Grand Central Station. As Schelling remarked, 'the chances [of a successful coordinated solution] are ever so much better than the bare logic of abstract random possibilities would suggest' (p. 57). The joint goal can require different actions from each party, as when during a telephone conversation the line is cut, and each party must independently but jointly decide who will put the receiver down to enable a reconnection. How coordination is achieved so reliably against such overwhelming odds remains I think a mystery; but both Schelling and later commentators have pointed out that it must have something to do with (a) the provision by the situation of a unique determinate clue, around which coordination can be achieved; and (b) some powerful property of the reflexive reasoning that inevitably comes into play: each must do what the other thinks that the other is likely to do. The two factors together, mutual salience and mutual computation of mutual salience, seem to be sufficient to turn a mere lottery into a near certainty.

Schelling (and later commentators like Schiffer 1972) was keen to point out that the computational problem posed by coordination is really very different from the formal properties exhibited by agonistic interaction, as explored in the mathematical theory of games (Luce and Raiffa 1957). In a zero-sum game (game of pure conflict), you can lay out the action–reaction sequences 'in extensive form' as a game tree or directed graph and calculate the relative utilities in advance of play. In contrast, in a cooperative game of pure coordination, each option of each player yields zero payoff unless it is matched by the coordinating option of the other player.[10] Both win if and only if each does what the other expects each to do; otherwise, both lose. In a zero-sum game, one's own preferences are clear in advance; in a coordination game, it doesn't much matter which action is taken as long as it matches the other's expectation. Zero-sum games can be reduced to relatively simple mathematics; but nobody knows if a mathematics for coordination games could even be formulated. Thus, calculating optimal behaviour in agonistic interaction is a far simpler computational problem than calculating coordination: strictly speaking, *Machiavellian intelligence* is child's play, a lower-order computational ability; *Humeian intelligence* (coordination through implicit contract) is the adult stuff.[11]



Curiously, though more than thirty years have elapsed since Schelling's work, there has been little empirical exploration of this striking kind of human ability to coordinate action through apparent 'mind-reading'. About the same time that Schelling was exploring these problems, Grice (1957) was devising his theory of (so-called) meaning, which is in fact a theory of communication which relies on intention-attribution. Although the theory has been around for thirty years, was subjected to thorough philosophical scrutiny twenty years ago, and continues to play an important role in theoretical work on communication, its relevance to empirical work has not generally been appreciated: it has appeared too complex, too intentional, too armchair philosophical to form a theoretical base for practical work. Recently, though, there has been a swing towards exploring its practical consequences in subjects as diverse as ethology (see, e.g., suggestions in Dennett 1988) and artificial intelligence (Perrault 1987; Cohen *et al.* 1990), not to mention linguistic pragmatics (Levinson 1983; Sperber and Wilson 1986) and the psychology of language (Clark 1992). But above all, it has stood the test of time, and remains a theory without a systematic rival of any consequence (see Avramides 1989 for recent commentary). The central idea is that communication is achieved when a recipient recognizes the special kind of intention with which a communicative act is produced. In one of many formulations it runs as follows:

*Grice's theory of 'meaning':*

S means p by x if:
S utters x
(a) intending to get H to think p
(b) intending H to recognize (a)
(c) intending (b) to be the reason for (a).

The point of the theory is that a communicative action is distinguished by its associated complex intention, which specifies that the 'signal' is a chunk of behaviour emitted solely (or at least largely) with the intention

of having its background intention recognized. (In contrast, many of our behaviours have an instrumental intentional background, where intention-recognition plays no part, as when I reach out to grasp a glass of water.)

Grice's theory of communication needs to be placed in the context of a general 'intentionalism', the view that any kind of interaction involves an attribution of meaning or intention to the other; it is discerning a chunk of behaviour as an action, that is, as a bundle of linked intention and behaviour, that is the prerequisite to response. The response in turn relies on the other's ability to read the intention or meaning from the behaviour. Of course there is usually a variety of ancillary information available to aid and abet this intention-attribution – preeminent sources being perhaps social roles which act to stereotype intentions (as E. Goody (1978a and this volume) suggests) and sequential patterns in interaction (see below). But producing behaviour in such a way that its intentional background is perspicuous requires a model of the other's ability to so recognize a behaviour 'x' as an expression of intention 'p'. Thus we enter the peculiar mirror world of reflexive intentions, now happily occupied by philosophers, computational logicians, theoretical psychologists and others. What distinguishes a Gricean reflexive intention from other kinds of complex reflexive intention is that the communicator's goal or intention is achieved simply by being perceived: recognition exhausts or realizes the intention.

There have been various attempts to marry Schelling's observations with Grice's, mainly with the aim of giving a philosophical account of how linguistic conventions may arise (and thus provocatively raising the possibility of reducing the concept of meaning entirely to psychological concepts (see, e.g., D. Lewis 1969; Schiffer 1972; Avramides 1989)). But there is a more direct and interesting application of Schelling's ideas to Grice's (Levinson 1985). For the obvious problem raised by Grice's theory is: how on earth are communicative intentions recognized? The traditional answer is: by means of a linguistic code (see Ziff 1971 on Grice). But this turns out to be no explanation even for linguistic communication, as is explained in the next section, and certainly not for non-conventional, non-linguistic communication. The fact of the matter is that we can communicate with 'nonce' signals (Clark 1992: ch. 10). An alternative answer suggested by Schelling's problems is that we can choose a behavioural token that is mutually computable as having been issued with a specific communicative intention, using the same techniques that allow us to coordinate on a unique solution to one of his coordination problems.

Assuming temporarily that some such picture is correct, let us take stock of the computational consequences so far. We are already in deep water:



*1. Propositional attitudes*

Obviously, computations about other's intentions presuppose computations over propositions embedded under propositional attitudes. As is well known, these are 'opaque contexts', contexts where Leibniz's law of the substitution of referring terms *salva veritate*, fails: we cannot assume from the assertion that *Esther believes the Chancellor of Cambridge University should be sacked* that she also believes that the Duke of Edinburgh should be sacked (she may think him competent, believe his title to be inalienable and certainly not realize he's the Chancellor). This well-known little conundrum is of course just the logical consequence of computations over other people's belief worlds (see, e.g., Fauconnier 1985), but it obviously raises difficult computational problems. There are a number of persistent logical paradoxes, like the Cretan Liar, which also plausibly have their roots here.[12]

*2. Mutual belief and infinite regress*

As mentioned, the Schelling problems seem to require that A and B, in order to coordinate, come to believe that some salient action is *mutually believed* to be the coordination point. But the notion of mutual belief seems to offer infinite regress: A must believe that B believes that A believes . . . *ad infinitum* . . . that p. This has attracted much attention, with philosophers (D. Lewis 1969; Schiffer 1972; Harman 1974; Avramides 1989), psychologists (Clark and Marshall 1981; Clark and Carlson 1982) and artificial intelligence (AI) workers (see, e.g., Allen 1983: 149) competing with different accounts each purporting to show how the regression can be circumvented. Nevertheless the threat of infinite regress has not endeared the idea of mutual knowledge to those interested in plausible models of psychological process. But the point of the Schelling experiments is that they demonstrate that people can indeed handle just this kind of reflexive reasoning.

*3. Gricean reflexive intentions and infinite regress*

As was early pointed out by Strawson, Grice's analysis alone might prove insufficient: one can produce counter-examples which satisfy the conditions but which intuitively are not cases of communication. These are cases where there is some higher intention of communicator S, not available to recipient H, and there is thus a discrepancy between the intention H is meant to discern and that which S actually has. This then threatens an infinite regress of conditions, with S intending that H should recognize that S intends that H should recognize . . . and so on.

Various proposals have been made to overcome this potential infinite regress. One is to ensure that all intentions are out in the open, as it were, if the behaviour is to count as genuinely communicative – Schiffer proposing for example that there be a condition of *mutual knowledge* that S has

uttered x with all the necessary intentions. But the notion of mutual knowledge must itself be cashed out as an infinite regress of the form 'S knows that H knows that S knows . . .', as we've just seen. Other solutions involve self-reflexive intentions (Harman 1974; see discussion in Avramides 1989: 58–9), or default inference rules relating *communicating p* to *believing that p* (Perrault 1988: 13).[13]

*4. Mutual salience*

We need not only a notion of salience, but a notion of 'natural salience', such that I can be sure, for an indefinite range of phenomena or scenarios, that what is salient for me is salient for you. This turns out to be crucial in ways I shall make clear (see also Schiffer 1972; for a variant suggestion see Sperber and Wilson 1986).

*5. Logic of action*

Clearly we need to compute the intentions that lie behind behaviours, if any kind of coordination is going to be achieved. That would seem to presuppose an understanding of the derivation of action from intention in our own planning and acting: one has to choose the means that will most effectively achieve the desired ends, while balancing incommensurable goals. As Aristotle argued, the logic of action is a distinct species of non-monotonic (defeasible) reasoning, a *practical reasoning* (PR) as it has been dubbed by philosophers. Von Wright (1971), Ross, Casteneda, Rescher and others have explored such systems, but there is still much to recommend the outlines of an Aristotelian system provided by Kenny (1966), which was developed into a formal system by Atlas and Levinson (1973) which we may call 'Kenny Logic' (see introduction in Brown and Levinson 1978: 69–70, 92–6). Kenny Logic has many interesting properties, like the fact that the deductive fallacy of 'affirming the consequent' is valid in this system, or the fact that if 'p' deductively implies 'q', then if an agent desires 'q' he'll desire 'p' (i.e., the logic of practical reasoning looks like 'backwards' logical implication, a fact that shows up in AI planning programs like Allen's – see the 'nested planning rule' in Allen (1983: 124)). But the relevant properties here are that Kenny Logic inferences are both *ampliative* and *defeasible*. They are ampliative because one may reason from a goal to a means that is more specific than is required to achieve the goal ('I'm thirsty and would rather not be so; here's a Coke; if I drink this Coke, I won't be thirsty; ergo I'll drink this Coke.'). They are defeasible because any valid inference from goal G1 to the desirability of action A1 will be abandoned if there is a conflicting goal G2 ('Coke is bad for my diet.'), from which the desirability of the negation of A1 can be derived. Such a logic of action must also explain how goals can be ranked, and means of achieving them differentially weighted, in such a way that the action performed may depend on the 'cost' of the means of achieving it.[14]



A logic of action is going to be a complex thing. However, all this turns out to be the least of the computational problems. Despite all the philosophical, logical and artificial intelligence work that lies behind all these ideas, there as been a fatal neglect of one problem. The Schelling-cum-Grice model of coordination and communication relies on *the recognition of intentions*: that is, the need to compute not only from intention to action (as in a logic of action or planning) but also in reverse as it were, from behaviour to the intention that lies behind it. It may seem that if we already have an account, in terms of a logic of practical reasoning, linking utterances or other actions with the goals that lie behind them, then all we now need to do is run the reasoning backwards, from the utterances or actions to the goals. Even logicians who should know a lot better talk as if intention-recognition is merely a matter of practical inference 'turned upside down' (as Von Wright (1971: 96) puts it in an uncharacteristic moment of incautiousness).

However, there is an overwhelming problem in equating understanding with 'upside down' practical inference, namely the very great difference between an actor-based account of actions (in terms of plans, goals and intentions) and an interpreter-based account (in terms of heuristics of various kinds). For the nature of logical inference in general, and practical reasoning in particular, is that *there can be no determinate way of inferring premises from conclusions*. Inferences are asymmetrical things. If I conclude from 'p and q' that 'p', you cannot, given the conclusion alone, know whether the premise was 'p and q', 'p and r', 'p and r and s', or 'q and (q p)', etc.: there are literally an infinite number of premises that would yield the appropriate conclusion. Simple though the point is, it establishes a fundamental asymmetry between actor-based accounts and interpreter-based accounts, between acting and understanding others' actions. There simply cannot be any computational solution to this problem, as so far described. The problem is intractable![15] Because the point is important let me put it in a more concrete way. Suppose I see you raise your arm outstretched in front of you: your doing this might be compatible, let us say, given the environmental possibilities, with the following intentions – waving off a fly, reaching for a glass, greeting an acquaintance, stretching, etc.[16] Even this set of descriptions is to 'cook the books': instead of 'reaching for a glass' why not go down a stage in specificity to 'extending an arm' or upwards to 'having a drink'? What we take to be a natural level of action description is anything but given (as philosophers from Anscombe to Davidson have been keen to point out). But then how do we decide what the hell you are doing, and what we should do in response (raise our hands too, do 'civil inattention', or

whatever is appropriate), all in the twinkling of an eye (say, 100 milliseconds)?[17] Going 'backwards' from the behaviour to the intention, at some appropriate level of specificity, is an absolute inferential miracle.[18]

### *Language, communication and interactional intelligence*

It is easy to imagine that the main role of language in the evolution of interactional intelligence is as an independent channel of information about others' plans and desires, which then makes coordinated interaction possible. That threatens to miss the point – *language didn't make interactional intelligence possible, it is interactional intelligence that made language possible as a means of communication.*[19]

Non-linguists may require a word of explanation. The model we used to have, both lay and expert (from Saussure to Shannon and Weaver), of the way that language works in communication goes something like this: we have a thought, we encode it in an expression, emit the encoded signal, the recipient decodes it at the other end, and thus recovers the identical thought. A moment's reflection will reveal that this picture is absurd. Consider the 'thing-a-me-jigg' phenomenon:

A: Where the hell's the whatdjacallit?
B: Behind the desk.

Just as in a crossword puzzle, the filled blank, *the whatdjacallit*, advertises itself to the recipient as a puzzle the recipient can solve. This works. In fact it works all the time – we don't say exactly what we mean. We don't have to and anyway we couldn't. For example, consider the relation *at* in

'The car is at the door.'

'The man is at the door.'

'He's at his desk/at University/at work/at lunch/at the telephone.'

There is no unified concept '*at*', except in some highly abstract way: we figure out the relation by thinking about the objects related and their stereotypical dispositions. Everything is amplified and specified through a complex mode of interactional reasoning.

The consequences of this kind of observation are rather far-reaching. Linguistic competence is not *sui generis* (at least not *this* part of it); it is not 'encapsulated' in Fodor's (1983) sense of a specialized, closed-off, module of mental processing. Semantic representations, or at least interpreted semantic representations, can't be the 'language of thought' – we think specifically, we talk generally.[20] I can't say what I mean in some absolute sense: I have to take into account what you will think I mean by it. One can't *encode* a proposition; all one can do is sketch the outlines, hoping the recipient will know how to turn the sketch into something more precise (if something more precise was intended). The slow realization of all this (Atlas 1989; Clark 1992; Levinson, in press; Sperber and Wilson 1986) portends a sea-change in the theory of language: linguistic mechanisms are deeply interpenetrated by interactive thinking.[21]



But if we can't say what we mean, how do we understand one another? When I say 'The coffee is in the cup', I don't have the same kind of IN-ness in mind as when I say 'The pencil is in the cup'. And when I said 'The coffee is in the cup', how come you didn't wonder: 'Does he mean the cup is full of beans, or granulated coffee, or the liquid stuff essential to academic life?' Nor for one moment, upon hearing 'The pencil is in the cup', are you likely to think of granulated pencils. Nor are you likely to worry that the pencil is more than half out of the cup, although on just those grounds we might expect a quarrel about the truth of 'The arrow is in the bull's eye'.

It is trying to understand mutual comprehension, given the paucity and generality of coded linguistic content that now preoccupies theoretical linguistic pragmatics (cf. Atlas 1989; Horn 1989; Sperber and Wilson 1986; Levinson 1989). We have made some progress in the last twenty years or so, by identifying heuristics that guide the reasoning process. I believe that the two cardinal achievements have been to identify two rather different kinds of heuristics. The one kind is a set of heuristics based on utterance-type, that is to say that the 'way of putting things' suggests a specific direction of interpretation. The other kind is provided by the intricate sequential expectations that are triggered by utterance and response in conversation.

To take these briefly in turn, the first kind of heuristic, which has been developed from seminal ideas of Paul Grice, in turn has a number of subtypes. These play off each other. For example, there seems to be an utterance-type heuristic that runs: 'normal expression indicates stereotypical relation'.[22] Consider expressions of the form *X is at Y*: when we say 'There's a man at the door', we have in mind a relation of proximity such that the man can reach the door-bell, say, and is facing it in expectation. But when we say 'Your taxi is at the door' it may be twenty feet away and its front not oriented to the front of the door. If your taxi was to nose its way in, the non-stereotypical event would warrant a non-normal description; while if the man waited twenty feet in front of the door, we might prefer another description, say, 'The man is standing some distance in front of the door'. That seems to be based on another heuristic: 'abnormal relation warrants abnormal/marked description'. The two heuristics together explain why 'It's possible to climb that mountain' and 'It's not impossible to climb that mountain' don't mean the same thing. A third heuristic runs: 'If an informationally richer

description applies, use it'. It's this that is responsible for the inference from 'Some of the Fellows of the College are lazy' to 'Not all of the Fellows of the College are lazy' – if you meant the stronger statement ('All of them are lazy') you should have used it.

In what follows I shall rely heavily on the importance of the inference to the stereotype. It's this that is responsible for such inferences as: 'The pencil is in the cup', suggesting 'The standard-type pencil (as opposed to, e.g., a propelling pencil or one with red lead) is projecting out of, but is supported by, the inside walls of the cup'. As we saw, we come to rather different conclusions from, 'The coffee is in the cup' (liquid rather than beans, fully within rather than projecting, etc.), or from, 'The key is in the lock' (projecting horizontally, not vertically). Some linguists will protest that these inferences are not pragmatic in nature but rather attributable to so-called *prototype semantics*. This I believe to be a rampant conceptual error, but regardless of that, it really makes little gross difference to the dimensions of the inferential problem: the particular relation intended by *in* for example still has to be inferred by reference to the things related.

The combination of these preferred interpretations of utterance-types can yield far-reaching enrichments of coded information. From 'Some of the nurses are not incompetent', one may infer that all the nurses are female (inference to stereotype), that not all of them are competent (informative strength), and that the remainder do not fully deserve the attribution of competence (marked description – the use of double negation). Or from 'If you wash the dishes, I'll give you 10 Deutsch marks' one may infer that if you don't, I won't (inference similar to inference to stereotype), that in any case I won't give you more than 10 DM (informational strength), etc. But I refer the reader to Horn (1989), Levinson (1983: ch. 3, 1987a, b) and Atlas (1989) for details. The point to grasp here is that *without* such inferential enrichments, what we say would tend towards the vacuous: not only do we talk generally, tautologically and elliptically (as in 'I'll be there in a while', 'If you manage, you manage', 'Could you please . . .?'), but also, as illustrated with the example of the relation *at*, even when we try to be precise we necessarily trade on suppositions our interlocutors must make.

The second kind of inferential enrichment that seems to me critical in language understanding is based upon the fact that, in the conversational mode that is the prototypical form of all languages use, speakers alternate, handing over to another party for response at the end of relatively short turns at speaking. And there's an expectation that responses are generally tied in close ways to what has gone before. As Sacks and Schegloff pointed out twenty years ago,[23] this makes it likely that if B responds to A in such a way that it is clear that B misconstrued what A said, there's a good opportunity provided in the third turn for A to correct, clarify or



elaborate. Thus recipients can be nudged along into what at least passes for understanding.

Take for example the following:

1. *From Terasaki (1976: 45)*

M: . . . Do you know who's going to that meeting?
K: Who?
M: I don't know!
K: o::h prob'ly Mr Murphy and Dad said prob'ly Mrs Timpte . . .

Here M asks a question of K; but K responds with a question ('Who?'). It is clear that K takes M's first utterance to be a prelude to an announcement, as in the canonical example that follows:

2. *From Terasaki (1976: 53)*

D: Y'wanna know who I got stoned with a few w(hh)weeks ago?
R: Who.
D: Mary Carter and her boy(hh) frie(hhh)nd.

There are systematic reasons why M's utterance in example 1 might be heard as the same kind of prelude or 'pre-announcement' as D's first utterance in example 2. But in any case, K got it wrong: M's utterance was not an offer to tell, conditional on K's not knowing the facts, but just a question as made clear in M's second, corrective, turn.

The power of such a system of feedback is well illustrated by the game of twenty questions: it's generally possible in just twenty question–answer pairs to guess what the other is secretly thinking of despite the fact that it might be anything under the sun.[24] In addition to such general corrective potentialities, we should add a large number of very detailed expectations about how particular sequences may run (like: question followed by answer, request followed by compliance followed by thanks, and so on). For a two-turn sequence A–B (like question–answer or offer–acceptance), each turn usually has rather restrictive specifications on form and content: the first turn because otherwise it will fail to be recognized as kicking off such a sequence, and the second because the first has been designed specifically to elicit it.

These sequential clues and constraints help to explain the rather astounding guesses that can be found in recorded conversations. In example 2 above, we saw an example of a sequential pattern that runs:[25]

A: Pre-announcement (request, offer, etc.)
((a turn that pre-figures what will come in third turn, conditional on B's signal to proceed))

B: Go-ahead
A: Announcement (request, offer, etc.)
B: Appreciation (acceptance, declination, etc.)

Mutual orientation to such patterns then helps to explain how a recipient can guess not only that something else is coming up, but that what will come is of a particular sort:

3. *Tape 170*

E: Hello I was wondering whether you were intending to go to Swanson's talk this afternoon
M: Not today I'm afraid I can't really make this one
E: Oh okay
M: You wanted someone to record it didn't you heh
E: Yeah heheh
M: Heheh no I'm sorry about that . . .

4. *From Terasaki (1976: 29)*

D: I-I-I- had something terrible t' tell you so uh
[
R: How terrible is it?
D: Uh, th- as worse it could be
(0.8)
R: W- y'mean Edna?
D: Uh yah
R: Whad she do, die?
D: Mm:mh

5. *From Terasaki (1976: 28)*

D: Didju hear the terrible news?
R: No. What?
D: Y'know your Grandpa Bill's brother Dan?
R: He died?
D: Yeah

Less dramatically, but more importantly and perennially, these sequential constraints help to explain how the often near-vacuous nature of what is actually 'coded' in conversation can carry so much meaning. In the following extract, for example, co-members of a band are haggling about how much they ought to practise together, and something as vacuous as 'Yeah I know but I mean' can serve to suggest that R's excuses for avoiding the next session really are not good enough:



6. *Tape 'Vicar' 144*

C: Yeah but I mean we'll be working all night
(1.0)
R: Uh [hh (I see)
C: er ( )(.) quite late
Well I mean it'- it's up to you I suppose
[
R: yeah
R: But I mean I've got the exam tomorrow so I can't
[
C: I mean I've
C: Yeah I know but I mean
(1.5)
R: Yeah alright yes =
C: = You understand what I mean
R: Yeah, do you want me to bring my guitar or not =
C: = Yeah

The limiting case is provided by the absence of speech altogether, which can alone be sufficient to engender detailed inferences, as in the following example where the speaker takes the absence of response to signify a clear negative answer:

7. *Tape: 'Oscillomink'*

C: So: u::m (0.2) I was wondering would you be in your office (0.63) on Monday (0.42)
by any cha:nce?
(1.86)
probably not

This example illustrates another important feature of conversational organization, namely that it has very precise temporal characteristics. Here, C has produced a pre-request in the form of a question, and here, as generally in English conversation, a pause of over half a second after such a question may be taken to indicate that the desired response cannot be easily produced. Due to such temporal characteristics, quite minute pauses can be most symbolic.

How does all this work? In the case of the utterance-type heuristics, it only works because speaker and recipient(s) agree that, other things being equal, there is a normal way to say things. That being so, a normal description can be taken to implicate that all the normal conditions apply, in all their empirical specificity: if I say 'John drove off, but he'd forgotten to loosen the hand brake', you envision a motor car and all the mechanical

consequences that such a failure of action would entail in such a mechanism. You know that I know that you will so imagine; you can therefore take me to be intending that you so imagine; and I can rely on you so imagining. You would be amazed if it later transpired that John drove off in a coach and four, or even a tractor! The same goes for the conversational sequences: you know that I know that you know that I expect an answer to my question within, say, 500 milliseconds; when you don't provide it, you know that I know that you have a problem – say, the desired answer can't be produced. Knowing that, I know from your silence that the answer is 'no', also of course that you are reluctant to give it, as you know that I know you should be ...

These examples, informally sketched, will suffice, I hope, to indicate the peculiar inferential richness that can be extracted by the combination of reflexive intentional reasoning and a handful of detailed mutual expectations. Conversational inferences have a number of very special properties: they are speedy, they are non-monotonic (the same premises can give different conclusions in different contexts), they are ampliative (you get more information out than went in) and they are subjectively *determinate*. The last point is important: when John says 'I'd like some water' we don't come away with a feeling that there's a 65 per cent probability that he had a glass of drinking water in mind, and a 35 per cent chance he was praying for rain; we come to a definite conclusion, which may of course turn out to be wrong (but then he'll tell us).[26]

In all this, conversational inferences are different from logical or monotonic inferences on the one hand, and inductive ones on the other. Inferring what is meant in conversation is much more like solving a slot in a crossword puzzle: such inferences have the rather special property of having been *designed* to be solved and the clues have been designed to be just sufficient to yield such a determinate solution. We might dub this central feature of language understanding the *whatdoyoucallit* property of language, in honour of the magical efficacy of that phrase.

Let me sum up these remarks about language so far. Linguistic communication is fundamentally parasitic on the kind of reasoning about others' intentions that Schelling and Grice have drawn attention to: no-one says what they mean, and indeed they couldn't – the specificity and detail of ordinary communicated contents lies beyond the capabilities of the linguistic channel: speech is a much too slow and semantically undifferentiated medium to fill that role alone. But the study of linguistic pragmatics reveals that there are detailed ways in which such specific content can be suggested – by relying on some simple heuristics about the 'normal way of putting things' on the one hand, and the feedback potential and sequential constraints of conversational exchange on the other. The astounding speed of conversational inference is something



that should also be noted; these are inferences clearly made well before responses can be composed, yet responses are, at least a third of the time, separated by less than 200 milliseconds, and on-line testing shows pragmatic inferences already well under way immediately after the relevant word or expression.[27]

## *Conclusion to section I*

We are now in a position to try and unravel the mystery. Recollect we concluded that the computational problem posed by Gricean communication or Schelling games looks simply intractable, largely because a system of inference from intention to behaviour tells us nothing about how to compute the reverse inference from behaviour to intent. And yet we routinely manage these things. The pragmatic heuristics may give us the clue to a solution. The inference to communicative intention from overt behaviour is so constrained by these heuristics or expectations that it is possible to select a unique path from within the interminable possible teleological explanations for the behaviour.

For example, if you know that I know that you know that, for principled and general reasons, a pause after a question seeking a 'yes'-answer will suggest a reluctance to provide it, then you know that your pause will lead me to think that you intended that I think that the answer is 'no'. Both knowing this, we both know that if you don't do something to correct the impression, then I'll feel sure that you wanted me so to think. Thus even the absence of a behaviour may be sufficient to yield the determinate attribution of an intention.

Or, you say: 'Put some bread and butter on the plates'. What do you intend? The stereotypical dispositions of course – not, for instance, a well-buttered plate, or bread on half the plates and butter on the other half. Even if there are only two plates, you're likely (in England anyway) to end up with buttered bread, not a plate of bread and a plate of butter. I don't need to ask you what you intended; I know that you know that we'll both be oriented to the heuristic authorizing inferences to the stereotype; so both you and I know that if you want something other than the usual, you'll have to make warning noises. You haven't; so the probability that what you wanted was buttered bread is now, for all current purposes, a dead certainty.[28]

How might this generalize from linguistic interaction to other forms of social interaction? In all cases, intention-attribution will be crucial, and the actual chunk of behaviour will be insufficient evidence alone for the attribution of an intent. We can carry out mental simulations: I can ask myself 'What would I be intending in these circumstance were I not me but him?' But that won't necessarily help me decide whether the

outstretched arm is a greeting or a reach for the drink or the beginning of a swipe at a fly; there are too many possibilities, too unconstrained a mental life in the other. What we need, just like the linguistic cases, is some basis for default presumptions – it really doesn't matter whether these are wrong or right, arbitrary or well motivated, etc., because once the expectations are in place you will know that I will know that you will think that I will use them to attribute an intention to your action, and then you can go about encouraging or discouraging that presumption (by modulating the behavioural manifestation).

It is in this context that Esther Goody's suggestion that social roles may play a special role in interactional coordination is, I think, important (see also Schelling 1960: 92). The point is made rather well by an artificial intelligence program designed by Allen (1983), one of the first to draw attention to the need for intention-attribution in the design of intelligent responses. The problem was to program an artificial railway-station information clerk. People might come to 'him' with questions like 'When does the train to Windsor leave?' or just 'The train to Windsor?' He ought to answer '11.15 at platform 5': that is, he ought to reconstruct the intentions behind the often elliptical question, e.g. that the traveller intends to find the train to catch, and then provide all the information relevant to that goal. How did Allen solve our intractable intention-attribution problem? Just by presuming that clients would come to the clerk presuming that his role was to answer travel questions, and they'd come with just two of their own goals in mind – catching a train or meeting one. Thus, by guessing the goals in advance, the program could simulate the plan-generation that might have led the client to say what he did, see if he could find an intention chain that culminated in that observable output, and then assume the client had those nested goals which the simulation used to arrive at the output (the client's question). In short, a presumption of the rights and duties of each party to the transaction made it possible to run the practical reasoning forwards, instead of in the impossible direction, backwards.

In the same sort of way, social roles may play a crucial role in ascribing intentions to our co-interactants. Often they won't have the intentions so ascribed – but that doesn't matter: by setting up the expectational background, interactants will know equally that they will have to do something rather special to escape it.

It is tempting, and not altogether implausible, to go the whole hog: what else is Culture, one might ask, other than a set of heuristics for intention-attribution? That clearly encompasses language usage, social roles (as just argued), and a host of heuristics for the interpretation of mundane and artistic productions. And why else do we feel so at sea in an



alien culture? We may understand the coded content of verbal interaction and fail to understand the import, observe behaviour but fail to comprehend its wellsprings, see mumbo-jumbo where we know there must be sense, and so on. Any facile definition of 'Culture' with a capital C deserves, no doubt, a certain modicum of derision, but I can think of definitions deserving louder hoots.

To sum up: human interaction, and thus communication, depends on intention-ascription. Achieving this is a computational miracle: inference must be made way beyond the available data. It is an *abductive* process of hypothesis formation, yet it appears subjectively as fast and certain – the inferences seem determinate, though we are happy to revise them when forced to do so. The extraordinary thing is that it seems, for all practical purposes, to work most of the time.

The question is: how? The best answer that we seem to be able to give at the moment is to take the Schelling games as model: there is an extraordinary shift in our thinking when we start to act intending that our actions should be coordinated with – then we have to design our actions so that they are self-evidently perspicuous. The crucial ingredients are (1) computations over reflexive intentions and mutual beliefs, and (2) the ability to settle on identical heuristics, mutually shared, which will yield default presumptions of intent. Without the heuristics, such coordination would not be possible: we have to agree in advance, as it were, what the salient features of the situation are, what any 'reasonable man' would think such a behaviour betokened, what one would 'normally' mean by saying 'He's at the door', and so on.

This kind of thinking turns mere probabilities into near certainties. Example: I try to guess your social security number (chances near to zero). I try to guess the seven-figure number that you have secretly chosen in the hope that I can guess it (chances over 0.9).[29] We can beat the odds. Otherwise humans couldn't coordinate in interaction. But it only works because we think there is a determinate solution, which we only have to find, like in a crossword puzzle.

That is the peculiar kind of thinking intrinsic to interactional intelligence. If interactional intelligence was the root of human intelligence in general, which is the idea we are exploring, then we would expect to find 'spill-over' effects in other task domains. For example, when thinking about (non-human) 'nature', we might expect to find 'nature' treated as a crossword puzzle, designed (by super-human agency perhaps) to be decoded and understood. And when humans come to think about chance, they should fail rather miserably to come to grips with the absence of non-deterministic solutions, with the fact that apparent patterns are in fact random assemblages, that a chance exemplar is not custom-made to be an

exemplar, that a sample could be unrepresentative, etc. It seems that there is in fact rather a lot of evidence from cognitive psychology in that direction, which I now review.

## II Biases in human thinking: psychological studies

### *'Judgement under uncertainty': Tversky and Kahneman*

Tversky and Kahneman (1977; Kahneman *et al.* 1982) conducted a series of now classic experiments on judgement under uncertainty – in effect intuitive responses to probabilistic problems. They found that despite the overwhelming everyday evidence to support basic principles of probability, people tend to follow other principles that yield incorrect conclusions. For example, everyday reasoning seems to ignore (or only partially take into account) the following basic statistical principles: (1) the prior probability of outcomes, (2) the confidence attached to large samples, (3) the potential independence of properties, (4) the possible chance occurrence of an expected outcome, (5) regression towards the mean, and so on.

More concretely, the following examples may give a clearer idea of the kind of errors systematically repeated:

*1. Neglect of prior probability*
If subjects are told, e.g., that X is meek, shy and very tidy, they'll guess X to be a librarian rather than a farmer even when told that there are twenty times as many farmers as librarians.

*2. Neglect of sample size*
If subjects are asked what is the better evidence that an urn contains $\frac{2}{3}$ red balls and $\frac{1}{3}$ white balls: (a) a sample of 4 red and 1 white or (b) a sample of 12 red and 8 white, they favour the smaller sample with the stronger proportion (even though the odds are half as good).

*3. Gambler's fallacy*
Given a sequence of 'heads-heads-heads-heads' even professional gamblers often presume a 'tails' must now be almost certain (the 'gambler's fallacy' (Tversky and Kahneman 1977: 330)).

*4. Neglect of regression towards the mean*
Trainers of airplane pilots come repeatedly to the conclusion that punishing bad landings has a much more powerful effect than rewarding good ones – failing to take into account that by natural regression the chances are that a really good landing will be followed by a worse one, and a terrible one by a better one (Tversky and Kahneman 1977: 332; Kahneman *et al.* 1982: 67–8). It seems not to have been noted by sociologists that this simple failure of statistical reasoning might be responsible for the vast asymmetry in the size of our penal codes compared to our system of honours!



Kahneman and Tversky attribute the majority of these 'mistakes' to a systematic bias to 'representativeness', i.e., 'a representative sample is one in which the essential characteristics of the parent population are represented not only globally in the entire sample, but also locally in each of its parts' (Kahneman *et al.* 1982: 36). This mistaken assumption of a representative sample explains the neglect of sample size in everyday reasoning: a small hospital's obstetrics ward is felt to reflect the sex ratio of newborns just as accurately as a large one. It also explains the ignoring of base-rate probabilities: if asked what is the probability of Mr X being a librarian, when X has all the stereotypical properties of librarians, then the fact that X is a typical representative candidate overwhelms thoughts about the rarity of librarians in the population at large.

Representativeness seems also to offer an explanation for the consistent and rather astounding tendency for people to ignore the most basic law of probability, Bayesian conjunction, whereby the probability of a joint event of greater specificity cannot be more than the probability of one of them alone. Thus the probability of John being both an accountant and a jazz player cannot be more than the probability of John being a jazz player – but subjects told that John is a compulsive person with mathematical skills and no interest in the humanities, feel the conjunction is more probable (which at least includes the representative profession) than the single attribute of being a jazz player (which alone seems unrepresentative (*ibid.*: 92ff., 496)). Representatives may also explain the gambler's fallacy: the feeling that having lost three times in a row, one is bound to win next time – the feeling being based on the expectation that a short run of dice should exhibit the same randomness of pattern found in a much longer run.

In short, single facts or small samples overdetermine conclusions because they are not considered in the larger picture of likely distribution. Instead, the subject's focus is on typicality, even though typicality and probability can obviously part company dramatically (e.g., an adult male weight of 157.85lb is highly typical, but less likely than a rough untypical weight of around 135lb (*ibid.*: 89)). These mistakes are intriguing because, not only do they fool the statistically naive, but also (sometimes only in less transparent examples) dedicated statisticians. It seems therefore that 'the bias cannot be unlearned', 'since related biases, such as the gambler's fallacy, survive considerable contradictory evidence' (*ibid.*: 30).

Kahneman and Tversky detect other, partially related biases. For

example, there is a tendency to presume *causal relations*, and to find it easier to infer effects from causes than causes from effects (*ibid.*: 118). There is also a strong bias to what the authors call *availability* (Tversky and Kahneman 1977: 333ff.), i.e., to salience or ease of recall, so that if it is easy to think of instances, then the event type may be thought to be frequent. For example, people overestimate the number of words beginning with R over the number with R in third place, because it is relatively much easier to retrieve words beginning with a letter than having such a letter in third position (Kahneman *et al.* 1982: 166). And quickly made associations are often presumed to be accurate correlations, despite evidence to the contrary: if recurrent correlation is one source of mental association, it is nevertheless of course illegitimate to assume that all associations are based on correlations (Tversky and Kahneman 1977: 335). Availability is thus a matter of *focus*; people overestimate the importance of what is in focus and underestimate what is out of focus: preoccupied with winning the lottery or with the thought of an aircrash, they overestimate the probabilities of both. One aspect of this directly relevant to agonistic interaction is the tendency to underestimate the opponent (Kahneman *et al.* 1982: 177).

Kahneman and Tversky conclude: 'In his evaluation of evidence, man is apparently not a conservative Bayesian: he is not Bayesian at all' (*ibid.*: 46). If this is correct, one must wonder why; after all, we live in a world dominated by chance events. It is self-evident that in the period in which our genetic makeup was laid down the dependence on chance, uncertain events, from the success of the hunt or harvest to the health of the chief or leader, must have been much greater than in the cybernetically controlled western world of today. How could we afford mental biases in such non-adaptive directions?

Kahneman and Tversky themselves offer no speculations, but they do offer us some tantalizing clues as to how their observations might tie into the biases or tendencies that are prerequisites for human communication – the overdeterministic mode of thought typical of, and necessary for, interactional coordination. They too notice the 'illusion of validity', 'the unwarranted confidence' which 'persists even when the judge is aware of the factors that limit the accuracy of his predictions' (Tversky and Kahneman 1977: 331). One such connecting clue is the obvious relation between their notion of representativeness and the notion of prototypicality (Kahneman *et al.* 1982: 86–9), as that latter notion has been explored in linguistic categorization. A representative individual would indeed be a prototypical one, and thus a special case of a representative sample, which should represent in microcosm the population as a whole, and thus also mirror its variability.



> If the reliance on representativeness leads to systematic errors, why do people use this relation as a basis for prediction and judgement? ... Modern research on categorization (Mervis and Rosch 1981; Rosch 1978) suggests that conceptual knowledge is often organized and processed in terms of prototypes or representative examples. Consequently we find it easier to evaluate the representativeness of an instance to a class than to assess its conditional probability. (Kahneman *et al.* 1982: 89.)

Our earlier suggestion, recollect, was that prototypicality or stereotypicality plays an essential role in communication and action coordination by providing a salient coordination point to which parties to interaction can each be sure the other will attend. Hence I can assume that you will understand what 'The pencil is IN the cup' would stereotypically suggest – if it lies in the bottom in broken pieces, I'd better tell you so. Likewise, if you move aside at the door, deference may be presumed as the motive rather than, say, fear. Stereotypicality provides the heuristic for a solution to the intention-attribution problem: each can assume that the other will use this heuristic and will therefore act accordingly, thus giving a deterministic solution to an otherwise impossible problem.

Other clues for connecting Kahneman and Tversky's observations to interactional intelligence may be found in their remarks about intuitive patterns of randomness, salience ('availability'), causality, sequentiality and so on. First, people find it unintuitive that short highly patterned sequences could be random. For example (where H = heads, T = tails), that HHH could be random, let alone HTHTHT or HHHTTT is counter-intuitive (Kahneman *et al.* 1982: 37). For even randomness is expected to be 'representative', i.e., exhibited over short stretches as an unpatterned sequence, for which one could write, for example, no little generative grammar. The corollary is: we see design in randomness, think we can detect signals from outer space in stellar X-rays, suspect some doodles on archaeological artefacts to constitute an undiscovered code, detect hidden structures in Amazonian myths. If we are attuned to think that way, then that is perhaps further evidence for the biases of interactional intelligence: in the interactional arena, we must take all behaviour to be specifically designed to reveal its intentional source. Second, people seem to favour higher probabilities where a causal or teleological connection can be posited, which is as might be expected from an interest in the wellsprings of action (*ibid.*: ch. 8); and they attribute quasi-intentionality to random processes when they act as if such processes were self-correcting (*ibid.*: 24). Further, they believe that they can somehow exercise control over chance, as we can over our fellows, as when they prefer a lottery ticket they have selected over one given out (*ibid.*: 236). Third, salience or mental 'availability', which Kahneman and Tversky

construe as a separate bias, plays a special role in the solution to Schelling games and communicational coordination: when we manage to meet again in a foreign city after getting accidentally separated, we do so by each thinking where the other will go, mutually deciding that a particular location (e.g., the café we were last together in) will be to each of us the most salient meeting place. Finally, when we find significance in the pattern of coin tosses THTH or think that after TTT we must have an H, we exhibit a sheer preoccupation with sequential pattern, where sequential patterning was one of the essential heuristics that we listed as making communication possible.

Consider again our treatment of communication as a 'crossword puzzle': there are multiple constraints on the 'slot' in which a communicative action is fitted, and the communicative act itself is only a clue to its proper interpretation. But it is a determinative clue: once you have it, you have it – you don't generally have it to 65 per cent certainty or the like; for it's taken to have been designed to yield a single, determinate interpretation. It's also taken to have sequential implications of a determinative sort – other communicative acts should now be opened up. The kind of thinking required in communication is a mental search for a salient – for example, stereotypical – interpretation, the psychological prominence of which is the best guarantee that this interpretation is indeed the mutually intended one. All the biases that Kahneman and Tversky list:

1. determinism, overconfidence, representativeness;
2. prototypicality;
3. sequentiality;
4. the ascription of teleology (e.g., in the belief in self-correcting random sequences),

would seem to be relatable to a communicational mode of thought – on this hypothesis they would be the side-effects of an interactional intelligence.

Of course one radical possibility is that Kahneman and Tversky's results are entirely a byproduct of the communicational context in which the experiments were carried out. Instead of focusing on the tasks as real-world problems, perhaps the subjects see the experimental tasks as communicative crossword puzzles: the experimenter has given clues as to what he wants – the subject must guess the desired outcome which has been designed to be guessed, like a problem in the classroom. Recently Kahneman and Tversky (postscript to Kahneman *et al.* 1982: 502) themselves have come to see that Gricean implicatures may play a role in their results through biasing the experimental description. However, they continue to underestimate that possibility severely. For example, they fail to note (*ibid.*: 497–8) that the famous Wason four card problem[30]



is entirely explained by what linguists call 'conditional perfection', the Gricean conversational implicature from 'If' to 'If and only if'.[31]

*8. Wason four card problem*

*Instruction:* Given the rule 'If a card has a vowel on one side, it has an even number on the other', which of the following cards should you turn over to test whether it is correct: A, B, 4, 7?
*Correct Answer:* A and 7. (Since the rules states 'If it's an A, it's even', or symbolically: 'A = even' which implies by *modus tollens* 'not-even = not-A'.)
*Predominant answers:* A and 4.
*Gricean explanation:* 'A = even' implicates 'A = even', thus 'even = A' (i.e. saying 'If it's an A it's even' implicates 'If and only if it's an A, it's even', from which it follows 'If it's even it's an A'.)

Similarly, there may be implicatural reasons for the failure to operate the Bayesian conjunction rule (that the probability of A and B cannot exceed either the probability of A or the probability of B). Consider the following task:

*9. Written background detail (Kahneman* et al. 1982: 496)

Linda is 31, single, outspoken and very bright. She majored in Philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.
*Instruction:* Which is more likely:

A: Linda is a bank teller;
B: Linda is a bank teller who is active in the feminist movement?

*Correct answer*: A
*Predominant answers*: B

The Gricean explanation here would rely on the presumption of a cooperative experimenter, who has produced (just as is always expected in conversation) only the relevant facts in the background description. But if judgement A is correct, then the facts must be irrelevant; since *ex hypothesi* the facts *are* relevant, A cannot be correct, so the only given alternative (B) must be right.

However, in addition to neglecting the possibility of a Gricean explanation of their alleged biases in thinking, Kahneman and Tversky have also failed to note the formal similarity between the basis of Gricean implicatures and some of their biases – those that we have noted under the rubrics of salience (availability), prototypicality, representativeness, etc. There

are thus two possibilities: the entire Kahneman and Tversky research programme is vitiated by the failure to consider the biases introduced by Gricean heuristics due to the verbal nature of the task-setting, or there is a real non-communicative bias in thinking, but one which mimics Gricean patterns because communicative heuristics inhabit, as it were, the deeper reaches of our minds. Given the breadth of experimental data, and the fact that at least some of the findings do not allow Gricean explanations directly, the second possibility seems the more likely interpretation of the facts.

### *Dörner's planning and decision-making with complex systems*

Kahneman and Tversky's tasks are presented verbally, thus opening them up to a critique in terms of communicative bias rather than cognitive bias. They also perhaps suffer from a parlour-game quality that may introduce Schelling's game coordination reasoning irrelevant to the design of the tasks. But there are other lines of psychological research that tend in the same direction. We turn now to just one further example.

Dörner (1990) has been exploring how subjects try to cope with complex, dynamic systems, often with hidden interconnected variables, delayed responses, patterns not available on short-term inspection, etc. Such systems, he argues, form an important part of the human decision-making environment, with good examples being agricultural and ecological systems, politico-economic systems, industrial plants, and so on.

The kind of problems posed by Dörner's simulations seem typical both of complex natural systems (like ecologies, predator–prey relationships) and complex artificial systems (like economies, polities, armies, industrial plant and other cybernetic machines). Yet our failures to understand quite simple indirect causations can have quite dramatic consequences (as with the Chernobyl accident, where increasing the flow of cooling water indirectly caused the removal of graphite moderators (Reason 1987)). Just as Kahneman and Tversky's results show a failure to understand the most elementary aspects of our largely random world, it is again rather striking that humans seem ill-adapted to coping with such complex systems, the manipulation of which might have been expected to have co-evolved with human intelligence. Surely, one might argue, a hunting-gathering prehistory might have led us to have intellectual mastery over predator–prey and other ecological systems, whereas in fact, of course, our understanding here has proved lamentable. Indeed, we carry one prototypical complex system permanently around with ourselves, namely our bodies, and yet it is notable how little we naively understand the complex system in which we are thus imprisoned – for example, how is it possible that the relation between sepsis and lack of cleanliness had to await discovery in nineteenth-century Europe?[32]



In Dörner's paradigm of research, unlike the Kahneman and Tversky one, there is often no mathematical or other precise way to measure optimal behaviour – in even a simple ecological system the variables are too many and their interaction too complex. Instead, one can measure the relative success or failure of individuals against the performance of their peers, in computer simulations of the relevant domains.

What Dörner has found is that when performance is bad, this is attributable to a number of recurrent 'errors', for example:

1. Failure to cope with delayed responses or long-term time series: e.g., failure to regulate a thermostat with a time-delayed, damped oscillation pattern (1990: 20). Subjects had a great tendency to react to the immediate state of the system, without taking into account underlying trends.
2. There was a tendency to interdigitate action and analysis, instead of doing prolonged analysis first.
3. There was a tendency to not look back and check the current consequences of past actions or guide them further.
4. While subjects often focus too hard on a salient problem, thus neglecting parallel problems and side-effects of the focal problem, they nevertheless jump from problem to problem, often without seeing the problem through to the end.
5. As subjects fail to control a complex system, they often take increasingly ill-adaptive measures – they seek for continued confirmation of failing hypotheses, become entrenched in their thinking and less observant of changes in the system.

Dörner attributes many of these 'failures' to a relatively small number of mental tendencies (1990). First, there are apparently irrational[33] or emotive factors, which he believes can be attributed largely to the desire to 'guard one's feeling of competence', especially by avoiding retrospective examinations of failures. Hence one finds, especially as the loss of control over a system builds up, a failure to check consequences of past actions, a 'dogmatic entrenchment' of failed hypotheses, a fleeting attention paid to one problem after another – a kind of mental 'panic' accompanied by behavioural rigidity, all a kind of escape behaviour, as subjects avoid facing up to the facts of loss of control of the system.

A second major source of failure he ascribes to sheer overburdening of mental processing. Hence subjects tend to seek single central explanations, or to judge a single variable to be the one determinative factor ('central reduction' (1987: 21)); and they fail to pursue consolidated prior information gathering.

A third factor may be the sheer inability to perceive patterns distributed over time because of *forgetting*; hence the failure to perceive time series, exponential growth patterns, etc.[34] A final factor Dörner isolates is

overemphasis on the current problem, with consequent neglect of side-effects and long-term effects, which he attributes to an attentional mechanism.

Dörner's explanations of the human failure to come to terms with complex systems can be pushed one stage further by asking why the 'errors' of thinking that he detects should be there in the first place. By innuendo, like Kahneman and Tversky, these mental failings are held up as natural deficiencies, as it were, for which no explanation is required. Thus 'human beings are "creatures of space" not "creatures of time"' and hence they fail to 'see' time-series or exponential growth patterns (1987: 22). The overall explanation is given in terms of a mixture of processing deficiencies and irrational self-deception.

Instead, accepting Dörner's analysis for a moment, one might try to find an explanation of the failings in our hypothesized interactional intelligence. One of his factors, the protection of one's feelings of competence, is easy enough to relate to interactional concerns. Self-esteem is not generally won or lost by encounters with 'nature', but rather by encounters with the fellow members of our species. We care about admitting mistakes to ourselves because we care about admitting them to others.[35] Thus if the 'irrational' factors involved in poor performance in struggles with complex systems can be rightly attributed to this preservation of the self from its failures, then these would seem to be deeply tied into factors at the heart of social interaction (although not the ones we focused on in section I).

The other key factors, the processing deficiencies, have certain striking similarities with the Kahneman and Tversky findings. Kahneman and Tversky's 'representativeness' carries with it a tendency to ignore the larger distributional pattern, and to focus on local patterns as if they were truly representative of the larger picture. Dörner's findings about restricted information-collection, restricted planning and the entrenchment of old hypotheses closely echo the Kahneman and Tversky findings, including that same insistent confidence in transparently erroneous inferences. For example, Dörner has found subjects to become totally preoccupied with the most salient problem, just as Kahneman and Tversky found naive thinking about uncertainty to be partially determined by 'availability' – i.e., salience, recoverability and focus. The same remarks about a possible source for such a focus in interactional intelligence carry over to Dörner's work: intention-recovery relies on coordination on a solution to the interactional 'crossword puzzle' – there must be a mutual focus on a single interpretation, and so the interactant must be forever seeking the single, determinative key to the intention-recovery problem. To this we can therefore assimilate also the tendency to seek just one single all-explaining factor, the critical variable on which the whole



complex system is thought to hang. Interaction requires single-solution thinking; complex systems require multiple-solution thinking – but humans are only good at the former.

Recollect that we suggested that there are two critical kinds of heuristics that make intention-recovery possible. One is the kind that yields a default interpretation for an action – the kind exemplified in verbal interaction by heuristics that legitimize the assumption of stereotypical attributes, and so on. The other is the kind based on sequential information. This latter kind shows up in the Kahneman and Tversky material as a presumption of the highly structured nature of any short sequence – thus accounting for the gambler's error, and the converse effect of the refusal to consider a patterned sequence as possibly random. Interaction sequences display certain clear properties:

1. The last action in an action chain is usually the focus of response (but not always, e.g. after an embedded 'insertion sequence').
2. Human interaction sequences which form canonical linear structures are usually rather short – for example a pre-sequence of four turns is relatively long (as in A: 'Doing anything tonight?', B: 'No, why?', A: 'How about having a meal together?', B: 'Great idea'.).
3. Interaction sequences are generally single-stranded – that is, one doesn't run two or more equally prominent chains of interaction simultaneously (there are of course exceptions, as with stockbrokers simultaneously dealing on phone and floor, etc.).
4. Interaction chains are characterized by rapid turn-taking, with short (average 400 millisecond) intervals between them.[36]

In Dörner's material one can find perhaps echoes of each of these expectations. The echo of property (1) is the tendency to focus on just the matter in hand, while being prepared to rapidly switch to other concerns. Echoes of property (2) would be the failure to discern much longer time series – the action–reaction style of human interaction ill prepares us for a complex system that reacts suddenly and catastrophically after a long time delay (as when an environment suddenly degrades, or when our bodies react to an earlier ingestion or infection). It also ill prepares us for reactions with relatively small time delays, but just sufficient to be beyond the normal action–reaction time span, as the subjects found in their unsuccessful attempts to control a simple time-delayed thermostat.[37] Echoes of property (3) (single-stranded interaction chains) are Dörner's observations about the single-strandedness of thinking, the preoccupation with only one causal chain at a time. Echoes of property (4) can perhaps be found not only in the temporal properties of the human attention span, but also in the tendency, for example, to collect minimal

information, then act, collect, act and so on, hoping to learn from interdigitated action–reaction rather than exhaustive prior analysis. Given the possibilities of interactional feedback, that is the right way to operate in the interactional domain.

Dörner's work, unlike Kahneman and Tversky's, is concerned with human abilities to cope with dynamic systems as they react over time – it thus serves to bring out some of the possible biases in temporal thinking that are not illuminated by the Kahneman and Tversky paradigm. It provides therefore a different kind of ammunition for the protagonists of the primacy of interactional intelligence. Those protagonists would be in trouble if Dörner's findings had indicated that people are good, say, at dealing with time-delayed reactions; or find long-term oscillations easy to discern; or can easily cope with multiple strands of sequential events. But Dörner's findings are all comfortably in the right direction – towards the conclusion that humans are good at dealing with single-stranded teleological or causal chains, with immediate action–reaction expectations which require immediate attention or allow only a small 'push-down stack', four or five 'plates' deep.

Kahneman and Tversky's results are vulnerable to the charge that all the observed biases are introduced by the linguistic communication that sets the task for subjects. Dörner's non-verbal tasks where the subject wrestles with a computer simulation escape that charge. However, they are arguably vulnerable to a parallel charge in just the area of most interest, the temporal characteristics of human behaviour: perhaps in setting up the computer simulations we have unwittingly introduced properties of human–human interaction into the design of human–computer interaction. For example, a keyboard or other input device will typically control only one variable at a time, nor are commands normally set to act at remote time intervals or at fixed delays to govern future states. To introduce such a system of relations between human and machine would be 'un-natural' – the very structure of the machines and programs we make for humans to interact with reflects (often to the rather lowly limitations of the engineer's imagination) the temporal properties that we, as humans, find it comfortable to work with. Thus the very set of biases we seek to illuminate in an objective way by setting subjects tasks that have intrinsically good solutions, may in fact have been built into the structure of the tasks themselves.

## Conclusions

I have argued that intersubjectivity requires peculiar computational properties, which may then bias many aspects of human thinking. On the one hand, one finds the presumption of deterministic solutions, what one may call the 'crossword puzzle effect' (problems are treated as if they were *designed* to be solved): hence the presumption that patterns can't be random, exemplars are prototypical, samples are 'representative' and conclusions can be certain. On the other hand, one finds some evidence that attention and memory are geared to interaction tempo: humans presume single-stranded causal chains, respond (usually) to the immediately previous event, expect brief action-response intervals and very short sequential patterns.



The first group of biases can be plausibly related to the necessity of having mutual orientation to the kind of heuristics we discussed as essential to language-understanding, e.g., the kind that gives us the strong readings of a preposition like *at*, according to the *relata*. The second group of biases, of attention and memory, may be related to the sequential heuristics for the attribution of intent in interaction – when talking, we are mutually oriented to the potential for immediate correction, and to canonical sequences of certain kinds.

Without such an explanation, the kind of biases noted by Kahneman and Tversky on the one hand, and Dörner on the other, would be puzzling indeed from an evolutionary perspective. The ability to make objective estimates of probability would offer immediate adaptive advantage, e.g., to a hunter faced with a decision to go after one kind of game or another.[38] Likewise the ability to comprehend complex systems, whether natural (like our own bodies, or the ecologies we live in) or socio-political, ought to offer significant adaptive advantages. It seems reasonable to suppose that, instead, there must be some greater adaptive advantage to thinking in the ways we actually do, and my suggestion is that these biases are essential ingredients for intersubjective reasoning. The corollary would be that the main evolutionary pressures on our species have been intraspecific. That accords with at least the views collected in Byrne and Whiten (1988), who have urged us to substitute a 'Lord of the Flies' scenario for a 'Robinson Crusoe' scenario for human adaptation. To that view, the speculations in this chapter add, hopefully, a corrective: it is cooperative, mutual intersubjectivity that is the computationally complex task that we seem especially adapted to. Machiavellian intelligence merely exploits this underlying Humeian intelligence that makes intersubjectivity possible. One needs too to stress that it is this cooperative intersubjective background that makes language interpretation possible (as shown by the need for all those heuristics) – not, as non-semanticists may assume, language which makes intersubjectivity possible (although it obviously vastly increases its scope).

In this chapter, I have stressed two pervading characteristics of human thought – attribution of intentionality and overdeterminism – which may be directly related to interactional intelligence. For without that over-

determinism, we would never have the heuristics that make it possible to ascribe intentions to human behaviour. As Peirce and many more recent writers have been keen to emphasize, deduction and induction are relatively trivial human skills, of no great computational complexity; it is abduction or theory construction which is the outstanding characteristic of human intelligence.[39] Abduction is the leap of faith from data to the theory that explains it, just like the leap of imagination from observed behaviour to others' intentions. While most explicit human theories or abductions are wrong, our implicit ones about interactional others are mostly good enough for current purposes. Both, though, come with that striking element of overconfidence, overdeterminism (even when we know, as in the case of scientific theories, that the half-life of the theory is only a year or two). Which allows me to end on a paradox: were we to feel any confidence that the roots of abductive ability (and it's peculiar phenomenology of certainty) lay indeed in interactional intelligence, and thus any confidence at all in the thesis of this chapter, then we could ascribe that feeling entirely to the overdeterminism of interactional intelligence itself.

## Acknowledgements

This chapter owes much to Esther Goody, who provided the stimulus to crystallize these thoughts. My second important debt is to London Transport, since some of the ideas here transcribed arose out of a long conversation with Dietrich Dörner while perforce walking the streets of London to and from meetings of the Royal Society during the transport strike, June 1989 (see Dörner 1990). Other ideas in this paper were first tried out at a meeting of the British Psychological Association (Levinson 1985), and I thank various commentators there, especially David Good, L. Jonathan Cohen and Phil Johnson-Laird. I have had helpful comments on this chapter from Penny Brown, Dietrich Dörner, Esther Goody, and Alex Wearing, for which I am most grateful. Much further back, Esther Goody (1978a) first pointed out to me (and us anthropologists generally I suspect) the relevance of the study of social interaction for theories about the evolution of human intelligence.

## Notes

1 I see the notion of *interactional intelligence* contrasting in specifity with other related notions. *Social intelligence* (or *social cognition*), as used for example in Flavell and Ross (1981), is an altogether broader conception, including the apprehension of morality, dominance, friendship and appropriate social role and affect. The *Machiavellian intelligence* of Byrne and Whiten (1988) is also wider, encompassing social knowledge, problem-solving in a world of flexible and fickle social relations, and so on (pp. 50ff.). By *interactional intelligence*, I have in mind just and only the core ability to attribute intention to other agent's actions, communicative or otherwise, and to respond appropriately in interdigitated sequences of actions; and I want to emphasize particularly the computational intractability of intention-attribution. I take this ability to be the bedrock feature of all the other, wider concepts, as recognized clearly in Esther Goody's term *anticipatory interactive planning* (AIP), which differs from my notion, I think, only in breadth and emphasis. All of these modified uses of the term *intelligence*, Alex Wearing points out to me, refer to a faculty or ability, and not the inherently comparative notion that the unadorned noun refers to.

2 For the prey orientation of the owl auditory system, see Schöne (1984: 212). For the matching of speech signal and auditory system in humans see, e.g., Lafon (1968: 81, fig. 2). However, in the human case there is more than mere matching of frequency and amplitude between speech signal and auditory discrimination; there is also a special kind of neural processing that clocks in when speech sounds are heard (Lieberman and Blumstein 1988: 148ff.). There are also fairly clear patterns of matching between properties of speech and properties of short-term memory (see note 37). Unfortunately, there is no one locus where all these sorts of facts are laid out for non-specialists, although they are essential background to speculations about the evolution of language.

3 See, e.g., Kolb and Whishaw 1990: 237–41; for a wider-ranging popular account see Landau 1989.

4 If there is such an assemblage of abilities that we can call interactional intelligence, why has it been so neglected in the wide range of disciplines (from anthropology to neurology) that might have studied it? Presumably, partly because of the tendency to take for granted what humans are naturally good at. We do not cherish bipedalism in the same way that we celebrate our ability to do calculus. The corollary is that we value that which we are not very good at: dancing *au point*, calculating decontextualized syllogisms, democracy, etc. But there may be another reason for the neglect of the study of interaction, namely inhibition or repression. It is not only that (as every transcriber of a conversation knows) friendly interaction is, on minute inspection, replete with nasty little jabs. It's also that certain human skills only run fluidly out of conscious awareness. Just as it is awfully hard to drive when taking a driving test, walk in a straight line when arraigned on suspicion of drunken driving, or appreciate a symphony when trying too hard to appreciate it, so self-conscious interactants generally do themselves a disservice (see, e.g., Field 1955 [1934]). If so, the repressive mechanism that aids our daily interaction may be responsible for making us equally reluctant to look at it scientifically. (On the role of inhibition in controlling, e.g., our perceptual world of smells, see O. Sacks (1985: 151).)

5 Dietrich Dörner suggests *Homo interagens*.

6 J.Z. Young (1951: 3) reporting Lord Keynes's comments on looking through Newton's alchemical papers: 'Newton was not so much one of the first men of the age of reason as the last of the magicians. He seems to have thought of the universe as a riddle posed by God, which could be solved if one looked hard enough for the clues. Some of the clues were to be found in nature, others had been revealed in sacred and occult writings.'

7 On speech to the self, see Goffman (1978) and Levinson (1988).
8 However, I return briefly to one social aspect in the review of Dörner's work below.
9 See, e.g., Goffman 1981; Levinson 1983: ch. 6; Clark 1992.
10 Humphrey's (1976: 19) seminal paper on the social function of the intellect uses the zero-sum game as a model of the computational demands of social life which 'asks for a level of intelligence ... unparalleled in any other sphere of living'. My point is that zero-sum games *merely* require decision trees for different contingencies; coordination games require deep reflexive thinking about other minds, and constitute a much more demanding intellectual task. In Schelling's (1960: 96) words: 'In the pure coordination game, the player's objective is to make contact with the other player through some imaginative process of introspection, of searching for shared clues; in the minimax strategy of a zero-sum game ... one's whole objective is to avoid any meeting of minds.'
11 Hence a superficial objection to the terminology of Byrne and Whiten (1988). Actually of course what they intend is a Machiavellian intelligence superimposed on a Humeian one, i.e., the potential for an agonistic exploitation of a supposedly cooperative understanding (cf. their quote (p. vi) from Machiavelli: 'For a prince, then, it is not necessary to have all the [virtuous] qualities, but it is very necessary to appear to have them.'). Nevertheless, one can't help feeling that their ethology is pervaded by the very agonistic bias (vicious struggle for survival of the fittest) that underlies the very Robinson Crusoe model (man's mind against 'nature') which they are complaining about. We all know cooperation is harder than conflict; it is not so obvious that one reason is that it's computationally harder too. (By the way, the reference to Hume is to *A Treatise of Human Nature* (III, ii.2.) where reflexive reasoning about the benefits of mutual cooperation is supposed to underlie our tacit acceptance of conventions (see Schiffer 1972: 137ff.).
12 'One of themselves, even a prophet of their own, said, the Cretans are always liars' (St Paul's epistle to Titus 1, 12). If the Cretan prophet speaks truly, then what he says is false; if he speaks falsely, then what he says would truly characterize him, but must nevertheless be false. The quotational aspects of the paradox are usually abstracted away from in philosophical discussion.
13 Do humans really go through all this reasoning about what each thinks the other thinks, and if so to what depth? The answers seem to be 'yes' and 'indefinitely deep' respectively, as is most clearly revealed where asymmetrical beliefs at a deep level are the name of the game, as in military strategy, paranoia, fraud and the like. Consider the beginnings of recorded western military strategy: e.g., Hannibal beat Scipio at the battle of Trebbia by making his centre only look like the normal thick phalanx, drawing the troops onto the wings, so the centre would collapse and the wings wrap round. Next time round the Romans might expect the same strategy, so this time Hannibal might stack the centre for a central concentrated punch: Hannibal's thinking that Scipio's thinking that Hannibal is thinking that Scipio will suspect a weak centre; Hannibal's hoping that Scipio will think all that but *not also* that Hannibal thinks Scipio will therefore weaken his centre to reinforce the wings making the centre an obvious target. However, suppose Scipio (or his successors) figure all that out – then they will thicken up the centre. Best then to repeat the Trebbia formation, but to bow the centre out so that it *really* looks packed, but in fact is a hollow crescent, designed to crumble. So Hannibal thought and won the battle of Cannae by another pincer movement from the flanks (Connolly 1978). If early classical military strategy went that deep, how deep was the reflexive thinking that, e.g., Kennedy and Krushchev got into over the Berlin wall/Cuban-missile crisis of 1961–2 (Gelb 1986)? For depth in cooperative reflexive reasoning, consider, e.g., irony and double irony (see Penelope Brown's contribution to this volume (Chapter 7)).



14 See Cohen *et al.* (1990) for some recent ideas here.
15 This has not of course prevented computational attempts to circumvent the problem (see, e.g., Allen 1983; Perrault, 1987; Pollack 1986a, b; and papers (especially by Kautz and Pollack) in Cohen *et al.* 1990.)
16 Lest this seem too academic a possibility, an anecdote: the Germans are great hand-shakers; when we were living in Berlin, our Hausmeister, for example, descended on one, regardless of one's current preoccupations, to grasp the hand on first and last sighting of the day. But Germans more used to casual Anglo-Saxon ways curb the custom. Puzzled at first, I found myself inspecting every hand-jerk during greeting/parting moments as a possible candidate for a proferred hand, only to find it turn, more often than not, into a buttoning of the coat or a struggle with a sleeve!
17 Since conversational response can routinely fall within 200 milliseconds of the prior utterance, if one modestly ascribes half of that delay to planning of the response, then that leaves only the other half for comprehension, including intention- or plan-recognition, of the prior utterance.
18 One is struck too by how our abilities here are not greatly helped by ratiocinative leisure. For example, historians make a modest, and lawyers an immodest, living out of pondering on, and quarrelling about, intention-attribution.
19 In papers circulated prior to the conference behind this volume, Esther Goody argues that, although primate interactive intelligence presumably preceded the origins of language, it is language that has projected us beyond our primate counterparts by allowing the management and codification of social interaction. If one thinks about linguistic ability as a relatively encapsulated human skill, then its acquisition might be an explanation for our zoom into a sapient state. But if, as this section sketches, linguistic ability is *necessarily and essentially parasitic on highly evolved interactive reasoning*, then language is not the evolutionary rocket fuel; it's the rocket (see here Sperber and Wilson 1986). One must then accept a synergistic explanation: higher levels of AIP make higher levels of communication possible, but equally vice versa.
20 If 'the language of thought' is rather independent of 'the language of communication', then I don't see the latter playing the crucial role in internal representation of AIP that E. Goody hypothesizes. Alex Wearing points out to me that the phenomenon of 'gist memory' might argue against my aphorism – thoughts bleached by time may not be so specific. But, at least when we communicate about our immediate environs the aphorism would seem to hold good.

21 Those who follow Chomsky in thinking of a core linguistic ability as a highly specialized, innate mental module, must now exclude semantics from that domain. But many of us think that central aspects of syntax too show the stigmata of interactive reasoning (see Levinson 1987b, 1991).

22 In what follows I simplify drastically from a complex, intricate, clockwork series of mechanisms (see Levinsom 1983: ch. 3; 1987b; 1991; forthcoming). For an alternative version, see Horn 1989.

23 See, e.g., Sacks *et al.* 1974; for further references see Levinson 1983: ch. 6.

24 This trick, though, may rely on something beyond the simple mathematics of set partition, e.g., the idea of the uniquely salient solution that lies behind Schelling's games of coordination.

25 It was striking that in the conference at which this paper was delivered, Drew, Streeck and myself all produced pre-sequences as prototypical examples of interactive planning. It then struck us that such four-turn ('pre-sequence') sequences are perhaps the longest canonical sequences observable in normal conversation, barring the 'rituals' of greeting and parting. This is surely striking, especially when one considers that in situations of asymmetric power and authority (of a kind frequent enough in human societies) one might expect 'superiors' to be able to impose their multi-staged 'plans' on 'inferiors'. (Indeed, such three- or four- stage planning hardly counts as a major intellectual achievement for *Homo sapiens* – Haimoff (n.d.) arguing that gibbon calls exhibit pre-sequential structure.) Instead of forward imposition of structure, what one finds in conversation is a *robust contingency*: no-one, almost regardless of status or rank, seems able to guarantee what will happen beyond the turn after next! I think a good case could be made that such turn-by-turn contingency argues for a fundamentally egalitarian state in the Garden of Eden: we are as a species adjusted to adjusting to others.

26 There is now a burgeoning literature on non-monotonic reasoning systems (see, e.g., Ginsberg 1987). But rather than viewing these developments as technical solutions to how conversational (and more generally interactional) inference might work, I view them more sceptically as systems that ape the results of inference under mutually assumed heuristics (see next section, and Levinson forthcoming: ch. 1). In short, conversational inference is the *Ur* form of default reasoning; default reasoning is not some peculiar unmotivated property of the human mind to be copied slavishly on machine models of intelligence – default reasoning is a mode of thinking that arises as a necessary solution to interactional coordination. It may then spill over to other domains of reasoning – that's the thesis of this chapter – but it is primarily motivated by the need to find a solution to intention-attribution.

27 See, e.g., Marslen-Wilson *et al.* 1982; or Tyler 1992.

28 Of course, we can enjoy the jokey qualities of such examples. But specialists in computational language understanding don't; they are plagued by just those 'silly' misconstruals we enjoy. They have no computationally tractable system of heuristics under reflexive intentional reasoning to rid themselves of these (to us) 'obvious' misconstruals. A machine can have a database full of semantic knowledge, and may be replete with knowledge about probable relations between things in the world, and still fail to find the 'obvious'. See, e.g., Herskovits (1986) on spatial relations like *at* and *in*.



29 Another example: 'I've lost my senile grandmother in the department store; I've got to think what she'll do, expecting her to wander blissfully on.' (Chances for a quick meeting: slim.) Versus: 'I've lost my wife in the department store: she'll be thinking where I'll be thinking she'll go.' (Chances for a quick meeting: good.)

30 See Johnson-Laird and Watson 1977a: ch. 9.

31 'Conditional perfection' was so christened by Geiss and Zwicky (1971). The inference is often subjectively very strong, as from 'If you pay me $5, I'll mow the lawn' to 'If you don't pay me $5, I won't mow the lawn'. In the Atlas–Levinson (1981) scheme of pragmatic inferences this is a generalized conversational implicature, attributable to the Principle of Informativeness, or Grice's second Maxim of Quantity.

32 Ethnomedicine might provide a rich area for the comparison of cultural modes of dealing with complex systems, the essential cross-cultural similarity of the body providing a natural control, as it were.

33 Dörner points out that such behaviour may be perfectly rational in the sense that there is a rational means–end relation; it is the apparent over-evaluation of the goal that inclines us to view such behaviour with analytic pity. But compare the importance attached to the preservation of 'face' in interaction (Brown and Levinson 1987).

34 Forgetting, on Dörner's analysis, is not (or not only) mere mechanical failure, as it were, but also the side-effect of abstraction, or pattern-determining processes (shades of Galton).

35 Brown and Levinson (1987) argue that the protection of a notion of self-esteem and the projection of esteem for alter, motivate much of the detailed patterning of social interaction. Thus we can ground in interaction Dörner's observation that 'In a certain sense, maintaining a positive self-image is the requirement for acting at all' (Dörner 1987: 36).

36 See Ervin-Tripp (1979), and references cited there, for temporal properties of turn-taking.

37 One might speculate, indeed, that the temporal characteristics of short-term memory have evolved just to cope with the short spans posed by the action–reaction interval, on the one hand, and the maximal conversational sequential pattern, on the other. There is, for example, a striking parallel between the maximum capacity of the short-lived buffer known as 'echoic memory' and MLU (or mean length of utterance in conversation). Or, as Alex Wearing has put it to me, the properties of short-term memory and limited information-processing capacity (which together necessitate frequent feedback) show how *Homo sapiens* is virtually hard-wired for high-frequency conversational turn-taking.

38 Such ratiocination is ethnographically real, as we experienced when working with Aboriginal people in Cape York, still much concerned with the success of the hunt or fishing expedition. It is not straightforward to work out the probabilities of whether the mullet will be running at Aylem beach and whether the water will be clear enough to spear such fish under conditions only half predictable from the base camp, or whether it might be better to head for more dependable but less rewarding line fishing off a mangrove swamp. That's the stuff and excitement of the hunter's life.

39 This is, of course, not the view of Piaget, who viewed the logico-mathematical as the apical intelligence, but it must now be a commonplace in the cognitive sciences. Computationally, bipedal locomotion is vastly more complex than calculus. What we can do 'without thinking' we devalue as not real thinking; hence our disregard for interactional intelligence. Curiously, though, some logico-mathematical tasks of the highest order are performed by 'idiot savants' who typically exhibit low IQs and gross interactional inabilities or autism (see O. Sacks 1985: ch. 23; and more scientifically, Howe 1989). They can calculate twelve-figure primes 'without thinking', a task for which there is no known algorithm.