Cognitive bias and the poetics of surprise

Vera Tobin

Language and Literature 2009 18: 155
DOI: 10.1177/0963947009105342

ARTICLE

Cognitive bias and the poetics of surprise

Vera Tobin, *Case Western Reserve University, USA*

Abstract

The 'curse of knowledge' is a pervasive cognitive bias that makes it very difficult for us accurately to imagine, once we know something, what it is like not to know it. This article analyzes examples drawn from both novels and films to demonstrate that this bias plays a substantial and previously unexamined role in narrative structure. I argue that narratives often take advantage of the curse of knowledge to solve an ongoing storytelling dilemma: how to engineer satisfying twists that genuinely surprise audiences but also avoid coming off as non-sequiturs or cheats. The curse of knowledge provides a useful mechanism to encourage readers to over-generalize propositions in predictable and reproducible ways, while making it likely that they will also agree, in retrospect, that these generalizations were mistaken. The same bias serves to enhance the impression, in hindsight, that the narrative's outcome was indeed possible to predict. Finally, building on Mental Spaces theory (Fauconnier, 1985, 1997) and simulation-based theories of language processing (e.g. Barsalou, 1999; MacWhinney, 2005), I argue that the curse of knowledge is an artifact of a more general cognitive shortcut that is implicated in features of 'correct' sentence interpretation such as presupposition projection as well as in phenomena traditionally described as curse-of-knowledge errors. This account unifies the discussed examples and helps to explain why certain devices, particularly unreliable narration, emerge so frequently as aids to narrative surprise.

Keywords: *cognitive bias; cognitive stylistics; curse of knowledge; mental spaces; narrative viewpoint; perspective-taking; presupposition; surprise; theory of mind; unreliability*

1 On reading and social cognition

Literary studies have lately shown a growing interest[1] in the relationship between fiction and theory of mind.[2] Work on this relationship has so far focused primarily on the fundamental observation that literature relies on and challenges the kind of thinking required to pass a false-belief test – a kind of psychological task that measures a person's ability to recognize that other people can hold beliefs that the person performing the task knows to be false (Wimmer and Perner, 1983; Leslie and Frith, 1988; Baron-Cohen, 1995). Lisa Zunshine (2003, 2006), for example, discusses the likelihood that texts like *Mrs Dalloway* are 'difficult' largely because their multiply-embedded representations of characters' beliefs make particularly taxing demands on this ability, and David Herman (2006) argues that narrative embedding has special aesthetic appeal because, among other things, it supports and recapitulates the everyday experience of thinking about other minds. The present work considers a different element of the relationship between literature and social cognition; namely, how the idiosyncratic mechanics of mental-state understanding are reflected in and exploited by narrative structure.

(Los Angeles, London, New Delhi, Singapore and Washington DC), Vol 18(2): 155–172
DOI: 10.1177/0963947009105342 www.sagepublications.com

Specifically, I will discuss consequences of a bias known as the 'curse of knowledge' (Camerer et al., 1989). Things that are highly salient in our own minds – the tune of the song we are tapping out on the table, our own beliefs and knowledge about the topic of conversation – tend to seep into our notions of what others know or have noticed, or our notions about the obvious characteristics of objects in our environment. Because our own intentions and privileged information about the intentions of others stand out in our own minds, it can be hard to remember how much less apparent they may be to others. Researchers who study these biases tend to regard them as a pernicious trap. I argue that they are, in fact, an artifact of a vital cognitive shortcut, and that they play a major and hitherto unexamined role in the construction and experience of narrative surprise.

Even quite young typically developing children have sophisticated skills of social cognition. By the time they are toddlers, they can express and respond to communicative intentions in impressively sophisticated ways, producing and understanding demands, assertions and questions that reflect a flexible implicit recognition that other people's intentions can differ from their own and may be influenced by their behavior. But young children are also substantially worse at keeping track of what other people believe than adults and older children are. Specifically, children younger than about four years old and older children with autism consistently fail false-belief tests. A common interpretation of these findings is that young children lack a 'theory of mind': that they cannot predict others' behavior in this kind of scenario because they do not yet have a concept of what a belief is, or, more generally, because they do not yet understand that other people have mental representations of any sort.

Even though adults do not have the trouble with false-belief tasks that three-year-olds do, they are far from perfect at keeping track of differences between others' beliefs and their own. A number of cognitive, social, and developmental psychologists (e.g. Birch and Bloom, 2003; Keysar et al., 2003; Birch, 2005) argue that these findings for adults and false-belief effects in children reflect a single fundamental bias in social cognition, one that persists across development, although it is more severe in children, perhaps because they have less inhibitory control than adults do (Leslie and Polizzi, 1998). This tendency is the aforementioned curse of knowledge, and there are a host of experimental results illustrating the degree to which people are biased by their own knowledge and awareness when attempting to appreciate a less-informed perspective.

For instance, people who already know the outcome of an event overestimate what other people know about it and how easily they should be able to predict that outcome (Fischhoff, 1975). They tend to believe that their internal states are far more transparent to others than they really are (Gilovich et al., 1998). They underestimate how ambiguous their own utterances can be and overestimate the helpfulness of their attempts at disambiguation (Keysar and Henley, 2002). When they know the solution to a puzzle, they overestimate how easy it will be for others to solve (Kelley and Jacoby, 1996).

For adults and children alike, then, 'theories of mind' are imperfect, and the consistent egocentric bias of social cognition affects the inferences people make in predictable ways. Meanwhile, some kinds of fiction have a certain degree of license, even a mandate, to fool their audiences. Cognitive biases like the curse of knowledge are a natural, ready-to-hand resource for complying with this mandate, and the shape of those biases dictates the shape of the narratives that exploit them.

2 Cursed surprises

It is a truism, if perhaps an overgeneralization, that stories that aim to entertain should be surprising.[3] It has also long been observed (e.g. Chatman, 1978) that the pleasure of narrative surprise nearly always depends on the complementary pleasure of predictability. An entirely unpredictable narrative element frequently qualifies not as a satisfying twist, but as an unsatisfying non sequitur. A narrative that would avoid this pitfall must include elements early on that are endowed with some significance that will only be visible later. However, this significance must, in retrospect at least, seem to have been available from the start, or, when the reader looks back, she or he will not be satisfied.

A satisfying surprise of this sort, which I will call a *narrative rug-pull*, must undermine expectations, while maintaining a sense that the undermining has all been done in a spirit of fair play. This is especially true in cases where the narrative revolves around a central puzzle, as in classic mystery stories, or where it features what Friedman (2006) calls a 'changeover' twist, in which an assumption about some very central element of the story, such as the identity of a central character, is overturned by a revelation late in the narrative. Specifically, this kind of surprise will be only successful if:

1 it is unexpected
2 it does not, in retrospect, conflict with the information otherwise presented
3 it inspires a significant reinterpretation of that information.

These constraints create a problem that has been articulated especially well by authors of the whodunit, though it is not unique to that genre. Dorothy Sayers explains the mystery-writer's form of the predicament as follows:

> The reader must be given every clue – but he must not be told, surely, all the detective's deductions, lest he should see the solution too far ahead. Worse still, supposing, even without the detective's help, he interprets all the clues accurately on his own account, what becomes of the surprise? How can we at the same time show the reader everything and yet legitimately obfuscate him as to its meaning? (Sayers, 1929: 97)

The challenge to the storyteller, then, is this: It's no good to leave out all the crucial information that might ruin the surprise. On the contrary, you must include it, but in such a way that you can be fairly sure most readers will overlook it until

you draw attention to it, later on. It has to be planted firmly enough that it will be remembered when the proper moment comes, but subtly enough that it will go unconsidered until then.

This kind of self-negating trickery is an order of magnitude thornier than a confidence game, because a con artist can be perfectly successful if the victim never recognizes the deceit. A satisfying narrative surprise, by contrast, requires that the author simultaneously present both a ruse and a delayed-action means by which the ruse will be destroyed. But how? Cognitive biases provide one possible answer.

Research on the curse of knowledge generally subscribes to the notion that the phenomenon is wholly regrettable. It may be a side effect of a generally efficient and desirable set of heuristics for social cognition, but those who study it tend to agree that it is an unfortunate one. Camerer, Loewenstein, and Weber (1989: 1246), for example, who first coined the term, see their research as 'grounds for pessimism' about people's ability to learn from either personal experience or the advice of others. It comes off even worse in pop psychology; Heath and Heath (2007: 19) describe the curse of knowledge as a 'villain' that 'consistently confounds our ability to create ideas'. In fact, however, narratives can turn the curse into a blessing, capitalizing on these predictable tendencies of the mind to generate aesthetically pleasing effects.

Consider the following sequence, which culminates about a third of the way through Graham Greene's *Brighton Rock* (1938). At the start of the novel, Charles Hale, in fear for his life, attaches himself to a friendly and phlegmatic woman named Ida Arnold. When they are parted for a moment on Brighton Pier, he disappears – not because he has run off, as she first thinks, but because he has died. The novel quickly makes it clear that Hale was murdered by members of a gang, on the orders of their young leader Pinkie, though we never learn exactly how the killing was done.

When an inquest determines that Hale's death was due to natural causes, Ida is dissatisfied. Correctly intuiting that the coroner's version of events is incorrect, she appoints herself the task of determining what really happened to Hale that day. The reader follows along with her as she sifts through evidence from a number of sources to find her way to a full narrative of the day's events. As she makes her inquiries, Ida draws a number of conclusions that accord with the privileged information the reader has been given about both Hale's and Pinkie's movements in the hours leading up to the murder.[4] The results of her investigations all point toward what the reader already knows: that Hale was murdered. And so the reader may feel fairly confident about what to expect when, at last, Ida sweeps into the police station, ready to confront the authorities with what she knows: Hale's death was ... '"Suicide," Ida said, "right under your noses" (Greene, 1993: 94).

The punchline works by playing on the curse of knowledge. If we are surprised, it is because we have been seduced into projecting aspects of our own knowledge onto Ida's partially represented perspective, only to have the rug pulled out from under us by the revelation that the two differ in a crucial respect.

3 Recognizing bias

The example from *Brighton Rock* turns out to bear a striking resemblance to the constructed narratives used in a set of classic experiments (Keysar, 1994) demonstrating that egocentric biases have a significant effect on the inferences that people make as they read narrative texts. These studies were designed to probe the question of how people keep track of what information counts as common ground between characters in a story they are reading – how much and what kind of attention they pay to the question of 'who knows what about whom', in the words of a related study (Lea et al., 1998). They are, in effect, false belief tests that even neurotypical adults often fail.

In one of these experiments, participants read a short narrative in which one character, June, recommends a restaurant to another character, Mark. After eating at the restaurant, Mark leaves a note for June saying, 'About that restaurant, it was marvelous, just marvelous.'

The vantage of the narration in this text is of the type that Genette (1980) calls 'nonfocalized', which is to say that the narrator provides more information than any single character knows, explaining with authority both June's and Mark's feelings about the restaurant. It is not, however, forthcoming in any explicit way about the characters' thoughts or intentions; instead, the text restricts itself to descriptions of their actions and their evaluations – for example, reporting that June 'really liked that place' or that (in Mark's experience) 'the food was unimpressive'.

The participants who read a version of the narrative in which they were told that Mark hates the restaurant tended to believe that June would take him to be sarcastic, while participants who read a version in which Mark loves the restaurant tended to believe that June would take him to be sincere, even though in both cases June has no access to the privileged information about Mark's intention. What's more, when asked why they believed what they did about June's plausible reaction, participants under both conditions tended to point not to their knowledge about Mark, but rather to features of Mark's utterance, such as the repetition of 'marvelous'.

Brighton Rock reports forthrightly that Ida's first suspicion is that Hale may have killed himself. She responds to the news that he is dead by asking 'Did he kill himself?' (Greene, 1993: 39), later muses 'If he did kill himself ... he did his job first' (43), and finally is certain that her Ouija board thinks so, too: 'Why, it's clear as clear ... Suici for suicide' (56). Surely the default interpretation, then, should be to assume that Ida is still entertaining this theory, unless we are told otherwise. As in the Mark and June story, however, *Brighton Rock* undermines this interpretation by presenting privileged information immediately before the reader is prompted to estimate how much of that information the relevant character has been able to work out for herself.

People are especially prone to curse of knowledge effects with respect to information that is highly salient in their awareness (Newton, 1990), and *Brighton Rock* accordingly goes to considerable lengths to ensure that the fact of the murder

is both highlighted and frequently reiterated. Its possibility is first raised in the opening sentence: 'Hale knew, before he had been in Brighton three hours, that they meant to murder him' (Greene, 1993: 9), and the rest of the first chapter dwells heavily on this fear. Hale hopes that '[t]hey hadn't the nerve to kill him in broad day before witnesses' (14), finds himself in a crowd where 'the people he was among seemed like a thick forest in which a native could arrange his poisoned ambush' (16), thinks 'of the thin wound and the sharp pain' (17) that may await him, knows that 'it was always easy to kill a lonely man at a railway station' (18), and ultimately realizes that '[t]his was real now: the boy, the razor cut, life going out with the blood in pain' (21).

Later, some 30 pages intervene between Ida's investigations and the moment where she feels ready to announce her official final conclusions to the police. Here, scenes featuring Pinkie and his colleagues reconfirm that Hale's death was indeed a murder: 'You act as if it was last year we killed Hale, not last week' (67); 'he jerked his narrow shoulders back at the memory that he'd killed his man' (84). By the time we switch back to Ida, and learn that she's been doing some thinking, the curse of knowledge trap has been set.

What is especially handy about exploiting the curse of knowledge for constructing narrative surprises in this way is that people both fall prey to it reliably *and* are also generally aware that it exists as a hazard (Epley et al., 2004), even if they are much quicker to recognize it as a danger for others than for themselves (Pronin et al., 2002). This state of affairs has two convenient results. First, it means that mere awareness of the pitfall is no guarantee against it, so that the technique can be effective even for audiences who have been taken in by similar gambits in the past. Second, it provides an avenue by which the audience can accept the misdirection as fair play.

Just as the participants in Keysar's experiment automatically project their knowledge of Mark's 'true' motivations into their estimation of what Jane will be able to guess, many readers can be counted upon to expect Ida to guess the whole of what they know. When it is revealed that she did not, the surprise is real, because it runs counter to expectation, but it also rings true, because people all have real-life experience with guarding against these kinds of mistakes and being proved wrong in just this direction.

I can produce a similarly reliable surprise in the classroom, by subjecting my students to a reenactment of the 'tapping study' conducted by Elizabeth Newton (1990). In Newton's study, adult participants were asked to tap the rhythm of a well-known song and then assess how likely it was that a listener would be able to identify the song. The discrepancy between the expectation of comprehensibility and reality was considerable: while tappers expected that about half the listeners would be able to identify the song, the actual success rate was only 3%.

When my students try the same experiment, two striking things happen. First, even though they have invariably *just been reading* about the fact that people over-attribute knowledge to others, and even report that they consciously moderated their guess, the students still greatly overestimate the number of classmates who will have identified the song. Second, when they discover just how few of their

classmates really did guess what song they had in mind, they always laugh – because they know exactly where they went wrong.

Something very similar happens in the aftermath of the culminating line in the *Brighton Rock* example. Because readers already have an operational model in which this tendency is understood as a shortcoming in their own inference processes, the mismatch between a surprised reader's expectations and the story facts as they are eventually reported can be understood as the reader's own 'fault'. Thus the fair play constraint is met and the surprise works.

4 Aligning perspective with an embedded viewpoint

In fact, there is good reason to believe that this kind of over-projection and revision together constitute a completely ordinary and even necessary component of meaning construction. The curse of knowledge is an artifact of an important, more general, cognitive shortcut: project what you know as far as you can. In many cases, the resulting 'cursed' thinking might more accurately be termed a curse of perspective than a curse of knowledge *per se*.

To see how this works, consider another example. At the beginning of Alfred Bester's novel *The Demolished Man* (1996 [1953]: 19–20), the anti-hero, Ben Reich, has sent a coded message to his archrival, urging a merger of their business interests. He receives a reply:

> The phone chimed one and then the automatic switched on. There was a quick chatter and tape began to stutter out of the recorder. Reich strode to the desk and examined it. The message was short and deadly:
>
> CODE TO REICH: REPLY WWHG.
>
> 'WWHG. "Offer refused." Refused! REFUSED! I knew it!' Reich shouted. 'All right, D'Courtney. If you won't let it be merger, then I'll make it murder.'

Later, however, the reader learns that 'WWHG' does not mean 'Offer refused'. It means 'Accepted', a fact confirmed by a list of codes that appears even earlier in the book, before the reader has any reason to think that any particular one of the codes might be noteworthy. It transpires that Reich killed D'Courtney for reasons that he could not admit even to himself, and that for this reason he 'was subconsciously compelled to misunderstand the message. He had to. He had to go on believing he murdered for money' (223). In the wake of this revelation, many apparently contradictory events in the plot make new sense. For example, the reason that the police have not been able to identify Reich's motives is that there is no evidence for the motives he believed that he had.

Something interesting has happened here, but it's not quite the same as what happened in *Brighton Rock*. There are some family resemblances, though: in both cases, the surprise works when the reader makes a mistaken assumption about the relationship between what a character knows and what the reader should take to

'really' be the case – in both cases, the assumption that the character's knowledge and the base level of the narrative match. But the origins of the propositions that are over-projected in these two cases are not the same. In the *Brighton Rock* surprise, the reader knows more than a character, and the surprise is that the character believes something different. In the case of *The Demolished Man*, we have too little information, rather than too much, and do not realize that we are underinformed. However, I would suggest that both surprises ultimately hinge on a general propensity for aligning one's own perspective with perspectives encoded in language. It seems that whenever we overpopulate possible viewpoints with what we know (or think we know), we are vulnerable to these sorts of rug-pulls.

Many accounts of linguistic pragmatics explain meaning construction in terms of mental representations. As far as I know, none of these accounts claims to model cognitive biases such as the curse of knowledge. However, Mental Spaces theory (Fauconnier, 1985, 1997; Cutrer, 1994) does have a theoretical apparatus to account for other ways that structure tends to flow from one mental representation to another. This apparatus offers an elegant solution to a number of classic problems in semantics, none of which are normally understood to have anything to do with egocentric biases in inferencing or shortcomings in reasoning about other minds. Nonetheless, I suggest that 'cursed' interpretations are in fact manifestations of the same principles.

The relationships among perspectives that a reader has to navigate in interpreting a narrative such as *The Demolished Man*, *Brighton Rock*, or the Mark and June story can be represented in terms of the mental spaces that she or he might construct in building up an interpretation of the text. Mental spaces are structured mental representations that can be reflected in and prompted by linguistic structure, 'partial structures that proliferate when we think and talk' (Fauconnier, 1997: 11).

These spaces are linked together in networks that are built up dynamically in working memory and can be stored in episodic memory. One mental space leads to another, and mental spaces can inherit structure from other spaces. These networks can be configured in a variety of different ways: event-chaining configurations that represent the conceptual structures invoked by the expression of tense, aspect, or causation; world-structuring configurations involved in the conceptualization of beliefs, possibilities, and stories; narrative configurations involved in understanding narrative embedding or free indirect discourse; frame-structuring configurations such as those involved in analogical thinking, and so on. Figure 1 shows the chain of (some) mental spaces and inherited structure involved in a narrative like the ones presented here.[5] Solid lines represent the trajectory of conceptual structures that are inherited 'correctly'; that is, it is not a mistake by any lights to assume that an author knows what her or his characters know. Dashed lines represent the propagation of information that could be attributed to the curse of knowledge.

Each of these spaces contains a representation of the knowledge and attitudes of some player in the narrative situation. These mental spaces, like any others, will

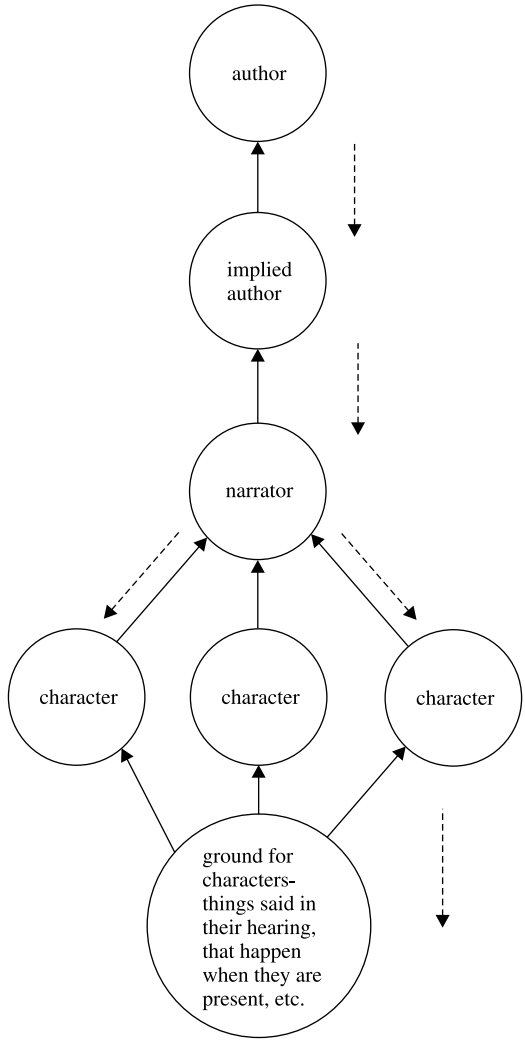

Figure 1 A narrative configuration of mental spaces

inherit some of their content from other spaces in the network. Propositions such as *The food at the restaurant Venezia is bad* or '*WWHG' means 'Offer Refused'* can thus propagate through the network.

According to Mental Spaces theory, conceptual structure is projected automatically into embedded representations via a principle called *space optimization*: 'relevant structure not explicitly contradicted is inherited within the child-space' (Fauconnier, 1997: 112). For example, the statement 'Sarah's so nice, I wish she were my sister' sets up two spaces: the base space of the speaker's reality, and a 'wish' space in which Sarah is the speaker's sister.

The sentence prompts us to make a specific counterfactual link between the base space and the wish space. In the base space, Sarah is not the speaker's sister. In the wish space, she is. There are many properties of Sarah in the base space that are blocked from being projected to the wish space because they are denied by this counterfactuality. For example, in the base space, Sarah and the speaker have different parents, do not share half of their DNA, and so on. We do not project these properties to the wish space. But other features of Sarah in the base space are unrelated to sisterhood, including her appearance, her age, and her niceness. These are projected to the wish space even though there is no explicit indication that they are the wished-for properties.

Structure is also automatically projected in the opposite direction, through a principle called *presupposition float*. It has long been observed that certain words and grammatical structures carry presuppositions, and that it is difficult to formulate a theory that will correctly predict which presuppositions of component clauses in a complex sentence will or will not survive for the sentence as a whole.[6] For example, both examples (1) and (2) presuppose that there is a king of France, thanks to the use of the definite description 'the King of France'.

(1) The King of France is bald.
(2) The King of France is not bald.

Other presupposition triggers include factive verbs such as *realize*, change of state verbs such as *stop*, clefts and pseudoclefts, non-restrictive relative clauses, and iteratives. Presuppositions typically hold up under negation, as illustrated by the King of France example, but it is more difficult to predict when presuppositions of embedded clauses will give rise to presuppositions of an entire sentence, or of a multi-sentence piece of discourse.

For example, the definite reference in a sentence like 'Mary wanted the King of France to visit her' does normally give rise to an interpretation that includes the presupposition that there is a king of France. But embedded presuppositions can in some circumstances be cancelled in a way that does not work when the reference occurs in a non-embedded clause, as in sentences like (1). Hence (3) is felicitous, while (4) is not.

(3) Mary wanted the King of France to visit her, but there is no King of France.
(4) The King of France is bald, but there is no King of France.

Fauconnier (1997: 61) explains the difference between these cases by the following rule: A presupposition floats up until it meets itself or its opposite. In other words, the default state of affairs is for certain kinds of structure to propagate through a network of mental spaces in all directions, only stopping when they run into a space where the contents actually contradict them.

The interpretation of 'Mary wanted the King of France to visit her' that includes the presupposition that there is a king of France is thus structurally parallel to the interpretation of *The Demolished Man* in which WWHG is taken to be the code for 'Refused'. The ultimate rug-pull in this narrative requires the reader to go through a specific sequence of cognitive construals: in the

first stage, the reader locates viewpoints in the narrative, recognizes that those viewpoints are embedded, detects some crucial propositions that derive from those viewpoints, and projects those propositions through the network. That projection is extensive and automatic. In the second stage, explicit prompts in the narrative lead the reader to revise the network connections so that the propositions do not project, but instead attach only to the embedded viewpoint in the other part of the narrative.

Similarly, Mental Spaces theory holds that structures that have been obtained implicitly, as through presupposition float, are liable to revision over the course of a piece of discourse. It is to be expected that conceptual structures will often be projected broadly through a network, only to be canceled later. 'Mary wanted the King of France to visit her' prompts the hearer to construct two mental spaces: a base space and a wish space. The wish space contains, among other things, the presupposed structure *the King of France exists*. By default, this structure is projected to the base space; but this projection can be cancelled after the fact. When the hearer is later told 'there is no King of France', most of the mental space structure set up in response to the first clause is maintained, but the projection of the presupposed structure into the base space is overridden on the basis of this stronger, explicit information.

Both presupposition projection and the projections underlying the 'cursed' readings of *Brighton Rock* and *The Demolished Man* seem to reflect a general disposition to align our perspectives by default with perspectives presented in a discourse. They also reflect an incrementalist approach to language processing, in which comprehenders begin interpretations as soon as possible and remain moderately open to shifts and reanalysis as they proceed (MacWhinney, 1977; Gernsbacher, 1990).

Recent work in neuroscience, cognitive psychology, and psycholinguistics on the role that simulated action and perception play in language understanding (e.g. Narayanan, 1997; Barsalou, 1999; Bergen et al., 2004) further supports the notion that language users often employ what MacWhinney (2005: 200) calls the 'enactive mode' of processing. In the enactive mode, a reader or hearer adopts a presented perspective as her own and simulates that perspective through the filter of her own experience. This conflation of perspectives is in keeping with the kind of default interpretation we see across curse of knowledge effects, presupposition projection, and narrative rug-pulls.

What's more, it seems that people are more likely to process utterances enactively when they occur in the context of an extended narrative discourse. 'The longer and more vivid our experiences', MacWhinney observes, 'the more they stimulate enactive processes in comprehension'. This is good news for storytellers.

5 Why unreliable narrators are so reliable

The more an embedded perspective dominates the narrative, the more the addressee will align her perspective with the embedded perspective as it is

presented. The more aligned these are, the more 'cursed' the addressee's inferences will be, and the more vulnerable he or she will be to a narrative rug-pull. This is why unreliable narrators remain an especially dependable resource for setting up this kind of surprise twist.

As observed, one of the striking features of these sorts of surprises is that it is perfectly possible to fall prey to them even once you've seen the same trick done somewhere else. Agatha Christie's *Who Killed Roger Ackroyd* (1926) scandalized readers at the time by presenting a murderer, Dr Sheppard, who is also the book's narrator. But knowing the twist to *Ackroyd* is no proof against being surprised afresh by Jim Thompson's *Pop 1280* (1964), in which Nick, the buffoonish small-town sheriff narrator, slowly reveals himself to be a ruthless and unhinged manipulator with a steel trap mind. This is in turn no defense against the surprises of Vladimir Nabokov's *Pale Fire* (1962) and its gentle accumulation of evidence that its narrator-through-annotation Charles Kinbote is either not who he claims to be or quite mad (or both), nor against the film *The Usual Suspects* (1995) and its revelation that the flashbacks narrated by Verbal Kint have been fabrications disguising his true identity as a criminal mastermind known as Keyser Soze.

Reception-oriented accounts of narrative unreliability tend to treat it as a process in which a reader asserts his detachment from and authority over a text. Thus Chatman (1978: 233) describes unreliable narration as a circumstance in which reader and author establish 'a secret communication' circumventing the perspective of the narrator. Yacobi (1981) describes it as the outcome of one of several strategies that readers use to resolve textual contradictions and impose an explanatory frame onto a narrative. Nünning (1999: 69) similarly sees it as an act of reconciling disturbing inconsistencies into a reassuringly consistent whole: '[t]he construction of an unreliable narrator can be seen as an interpretative strategy by which the reader naturalizes textual inconsistencies that might otherwise remain unassimilable'.

In other words, deciding that a narrator is unreliable is frequently taken to serve primarily as a mechanism for banishing anxieties produced by an ambiguous text. The interpretation is reassuring; it provides comfort and structure; it makes the reader complicit with, rather than a dupe of, the author. But the revelation that a narrator has been unreliable can also be a shocking (if pleasant) disruption to a previously untroubled reading experience.

The twist in *Roger Ackroyd* caused a genuine sensation at the time of its publication, even though similarly structured revelations were common enough in previous works, Edgar Allan Poe's 'The Tell-Tale Heart' (1843) being perhaps the most famous example. Indeed, a number of contemporary commentators took the position that, as Christie's contemporary Willard Huntington Wright (1927) put it, 'the trick played on the reader in *The Murder of Roger Ackroyd* is hardly a legitimate device of the detective-story writer'. However, Dorothy Sayers (1929), among others, disagreed: 'this opinion merely represents a natural resentment at having been ingeniously bamboozled. All the necessary data are given.'

This dispute depends on the ease with which the reader is able to take back the conflation of viewpoints after the fact. The surprise arises from structural

consequences of the fact that Sheppard is the narrator. His embedded perspective dominates the narrative, encouraging readers at every turn to align their own perspective with it. There is something especially unsettling about the revelation that a trusted narrator has been unreliable. It is so easy to succumb to the curse of knowledge in this way that it happens invisibly. Some readers felt not just fooled, but downright betrayed by the way Christie had disguised her most crucial clues.

Another reason that Christie's version of this twist was so effective and surprising, of course, was the way that it played on other conventions of the genre. The narrator of *Roger Ackroyd* fills a conventional role, that of the faithful sidekick exemplified by Dr John Watson, the friend and confidant of Arthur Conan Doyle's famous detective Sherlock Holmes. As the narrator of the Holmes stories, Watson serves as both the detective's assistant and the reader's proxy. He is clever enough to follow Holmes' reasoning when it is explained to him, but never so excessively brilliant that the explanation would be unnecessary. He is capable and faithful; if he were not, the perceptive Holmes would surely not rely on him.

Watson characters abound in classic detective fiction, from the prototype unnamed narrator in Poe's Dupin stories, to Watson himself, and Hercule Poirot's own usual sidekick Arthur Hastings. The Watson character serves his own conventional purpose: He gives the reader access to all the relevant clues, and to tantalizing hints about the detective's superior lines of deduction, without revealing too much about their conclusions until the proper time. This convention reinforces the bias effect; while one can often overcome the curse of knowledge given the opportunity to consciously reflect on the possibility that it might be a mistake, the conventional role of the Watson character helps to discourage readers from doing so.

6 Doubling down

Of course, readers are nothing if not resourceful, and there is plenty of room to construe 'inconsistent' or 'incoherent' texts in ways that preserve aesthetic pleasure. Not only are ambiguities pleasures in their own right, many readers also delight in constructing possible scenarios or readings in which these inconsistencies are explained away and naturalized. But this interpretive experience does not produce the rug-pull effect I have been describing here. A narrative that seems merely inconsistent may indeed make readers feel like the rug has been pulled out from under them, but on a different level – not intentionally, not pleasurably, but through a failure of the text.

In the end, curse of knowledge surprises succeed to the level that they provide a sufficient degree of plausible deniability to satisfy the reader. As long as crucial (mis)information can, in retrospect, be traced to a failure to appreciate possible discrepancies between represented viewpoints, it is safely available for reinterpretation. To the extent that the reader accepts the new interpretation, the misdirection qualifies as fair play.

Revisiting the original reading experience in light of the new information is part of the pleasure of the rug-pull. In the famously rug-pulling movie *The Sixth Sense* (1999), for example, the young protagonist, Cole, has the unique ability to see dead people, who walk unseen among the living. For most of the movie, he struggles with this second sight with the help of a sympathetic child psychologist, Malcolm Crowe, who is struggling with demons from his own past. The revelation that Dr Crowe is one of the ghosts that only Cole can see is the major surprise of the film, requiring naïve, first-time viewers to reassess many of the events of the film in ways that differ radically from their original apparent significance. To drive home both the surprise and the fair play, this revelation is followed by a rapid replay of the many moments when Crowe's ghostly qualities were on display, but (the movie presumes) overlooked, so that the intended viewer can appreciate both aspects of the surprise at once.

It is at this point that the curse of knowledge makes its second major contribution to the poetics of surprise. We have already seen that texts can use the curse of knowledge to guide readers into making certain predictable inferences about the fabula, or underlying 'facts', of a narrative. This predictability is a crucial condition for setting up a satisfying surprise, because if there is no reliable way to guess what readers have inferred, how can you know if your revelation will conflict with it? The second half of the trick is to ensure that the reader feels, as Sayers put it, that she 'has been given every clue' – that the narrative is internally consistent; that characters may have lied, but the author never did; that she *might* have guessed the surprise ahead of time, if only she had been more astute. We have seen that the curse of knowledge helps narratives to meet this requirement by virtue of the fact that people recognize this bias as a potential flaw in their own thinking. But the curse of knowledge also directly affects the way that people retrospectively assess the earlier parts of the narrative.

Paul Sheehan (2002: 13) suggests that 'a successful narrative configuration depends on the movement from contingency to inevitability'. It is the tendency of audiences, possessed as they are of human minds, to project that inevitability backward onto what once seemed contingent. Just as people have trouble discounting their own knowledge when they try to assess how other people will interpret a situation, they have trouble appreciating the way things once appeared to their own earlier selves. When people think about the past, known outcomes seem to have been more predictable or obvious than they actually were at the time. For instance, reading or hearing medical case studies – a discourse genre with many structural similarities to classic mystery stories – made doctors believe that the discussed diseases were easier to diagnose than they actually were (Dawson et al., 1988). Similarly, after having seen the flashback, many viewers of *The Sixth Sense* wondered how they could have failed to notice that Dr Crowe never had a sustained interaction with any character other than Cole. This tendency usefully reinforces the interpretive experience required for a satisfying rug-pull.

The rule for this effect is that the 'solution' or revelation should seem, in hindsight, to fit naturally with the information otherwise presented. Conveniently, our curse of knowledge bias encourages this very interpretation. Provided that the

revelation seems reasonable, consistent, and appropriate enough that the reader can accept it as a plausible outcome of or explanation for the previously narrated events, the curse of knowledge will enhance the effect, making the revelation seem retroactively obvious and inevitable. It can make a good-enough fit feel exactly right.

7 Conclusions

The narratives discussed in this article are structured the way they are because our minds work the way they do. Narrative texts offer many opportunities to present embedded perspectives, such as the reported beliefs of non-narrator characters, or the viewpoint of a self-conscious narrator. The more extended the readers' exposure to an embedded perspective, the more likely they are to align their own viewpoint with that embedded view, and then to fall prey to 'cursed' thinking, failing to discount this additional information when imagining what others think. Texts can take advantage of this tendency to surprise readers with information that contradicts the over-generalized propositions. These surprises qualify as satisfying twists because they both are unexpected and, in retrospect, feel consistent with the information otherwise presented in the text. The curse of knowledge can further reinforce this effect, by making readers more likely to see the revelation as inevitable once it has been made. Conventions that originally evolved to fulfill other goals, such as verisimilitude, allowing the reader to be privy to a detective's processes but not his conclusions, or circumventing tedious exposition, can be exploited to further facilitate these effects.

Humans are endowed with a remarkable ability to adopt the perspective of other people, and the dynamics of perspective taking are crucial to the pragmatics of both conversational and literary discourse. However, these abilities do not make communication a process of perfect coordination, in which people flawlessly model what other people intend, believe, and know. Instead, communication, including literary discourse, is riddled with egocentric biases such as those arising from the curse of knowledge.

The curse of knowledge and related phenomena are an artifact of cognitive shortcuts that are in fact crucial for using language and understanding others, and which make it possible for us to perform the enormous amount of pragmatic inferencing required for ordinary communication. The apparent trap of these biases also provides rich material for sophisticated narrative effects, and provides a solution for bridging apparently contradictory requirements for aesthetic satisfaction in the construction of narrative surprise.

Notes

1 In addition to work on this relationship in print, the subject has also been a popular one at conferences, and an entire conference dedicated to 'Theory of Mind and Literature' convened at Purdue University in 2007.

2 Broadly speaking, the term 'theory of mind' is used in the cognitive sciences to refer to any ability to understand that other people have minds like one's own, with thoughts, beliefs, desires, and intentions that may be different from one's own mental states, and to the ability to hypothesize accurately about what these mental states might be. However, there is a great deal of debate about what such a 'theory of mind' consists of. Some researchers have argued for a so-called 'theory theory' (e.g. Morton, 1980; Gopnik and Wellman, 1994; Gopnik and Meltzoff, 1997), in which understanding others' minds involves an explicit representation of other people's mental states, based on an actual, if unconscious, theory of how these other minds operate. Others (e.g. Goldman, 1992; Harris, 1992) have argued for a simulation account, in which our ability to understand others is based in a process of simulating others' thoughts and feelings, rather than theorizing about them. Because of these theoretical disagreements, many prefer the terms 'concept of mind' or 'social cognition' to 'theory of mind', while some continue to use the latter as a theoretically neutral term of art.

3 See Sternberg (1978, 2003; *inter alia*) for influential articulations of the role that surprise plays in a variety of narrative genres. Brewer and Lichtenstein (1982) provide experimental support for some of his claims about surprise, information management, and story enjoyment. For an extended discussion of these narrative imperatives from a prescriptive standpoint, see especially chapters 10 and 16 of the widely read and cited 'screenwriter's bible', *Story: Substance, Structure, Style, and the Principles of Screenwriting* (McKee, 1997).

4 For example, Ida realizes that the discrepancy between the name Hale gave her, Fred, and his official name, Charles, arose because 'a man always has a different name for strangers ... and a man don't have a different name for every girl' (Greene, 1993: 41). This is exactly right, as the reader knows from the direct insight into Hale's own thoughts provided in the previous chapter. Ida's conclusions similarly accord perfectly with what the reader already knows when she determines that a crucial witness saw not Hale, but someone else pretending to be Hale, on the day that he died. Finally, her fundamental conviction that Hale died of something less natural than a heart attack is doubly confirmed by what the reader has seen of both Hale's and Pinkie's thoughts and actions.

5 Here I am not only grossly oversimplifying the narrative structure, but also leaving out almost everything about the actual and represented discourse context. This is in part to avoid committing myself to positions on which I would prefer to remain agnostic (for example, whether it is correct to consider author and reader co-discourse participants in a language event) and partly to lay bare the relationship to existing accounts of presupposition within Mental Spaces theory. For a much richer model of these relationships, see Werth, 1999.

6 See Levinson, 1983: 191–225 and Beaver, 2001: 101–34 for reviews of the extensive philosophy and linguistics literature on this problem.

References

Baron-Cohen, S. (1995) *Mindblindness: An Essay on Autism and Theory of Mind*. Cambridge, MA: MIT Press.

Barsalou, L.W. (1999) 'Perceptual Symbol Systems', *Behavioral and Brain Sciences* 22(4): 577–609.

Beaver, D.I. (2001) *Presupposition and Assertion in Dynamic Semantics*. Chicago, IL: Center for the Study of Language and Information.

Bergen, B., Chang, N. and Narayan, S. (2004) 'Simulated Action in an Embodied Construction Grammar', in K. Forbus, D. Gentner and T. Regier (eds) *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society*, pp. 108–13. Mahwah, NJ: Lawrence Erlbaum.

Bester, A. (1996 [1953]) *The Demolished Man*. New York: Vintage.

Birch, S.A.J. (2005) 'When Knowledge is a Curse: Children's and Adults' Reasoning About Mental States', *Current Directions in Psychological Science* 14(1): 25–9.

Birch, S.A.J. and Bloom, P. (2003) 'Children are Cursed: An Asymmetric Bias in Mental-State Attribution', *Psychological Science* 14(3): 283–6.

Brewer, W.F., and Lichtenstein, E.H. (1982) 'Stories Are to Entertain: A Structural-Affect Theory of Stories', *Journal of Pragmatics* 6(5/6): 473–86.
Camerer, C.F., Loewenstein, G.F. and Weber, M. (1989) 'The Curse of Knowledge in Economic Settings: An Experimental Analysis', *Journal of Marketing* 53(5): 1–20.
Chatman, S. (1978) *Story and Discourse*. Ithaca, NY: Cornell University Press.
Christie, A. (1926) *The Murder of Roger Ackroyd*. London: William Collins & Sons.
Cutrer, M. (1994) 'Time and Tense in Narrative and in Everyday Language', unpublished PhD dissertation, University of California, San Diego.
Dawson, N.V., Arkes, H.R., Siciliano, C., Blinkhorn, R., Lakshmanan, M. and Petrelli M. (1988) 'Hindsight Bias: An Impediment to Accurate Probability Estimation in Clinicopathologic Conferences', *Medical Decision Making* 8(4): 259–64.
Epley, N., Keysar, B., Van Boven, L. and Gilovich, T. (2004) 'Perspective Taking as Egocentric Anchoring and Adjustment', *Journal of Personality & Social Psychology* 87(3): 327–39.
Fauconnier, G. (1985) *Mental Spaces*. New York: Cambridge University Press.
Fauconnier, G. (1997) *Mappings in Thought and Language*. Cambridge: Cambridge University Press.
Fischhoff, B. (1975) 'Hindsight Does Not Equal Foresight: The Effect of Outcome Knowledge on Judgment Under Uncertainty', *Journal of Experimental Psychology: Human Perception and Performance* 1(3): 288–99.
Friedman, S.A. (2006) 'Cloaked Classification: The Misdirection Film and Generic Duplicity', *Journal of Film and Video* 58(4): 16–28.
Genette, J. (1980) *Narrative Discourse* (trans. J. Lewin). Ithaca, NY: Cornell University Press
Gernsbacher, M.A. (1990) *Language Comprehension as Structure Building*. Hillsdale, NJ: Lawrence Erlbaum.
Gilovich, T., Savitsky, K. and Medvec, V.H. (1998) 'The Illusion of Transparency: Biased Assessments of Others' Ability to Read our Emotional States', *Journal of Personality and Social Psychology* 75(2): 332–46.
Goldman, A.I. (1992) 'In Defense of the Simulation Theory', *Mind and Language* 7: 104–19.
Gopnik, A. and Meltzoff, A.N. (1997) *Words, Thoughts, and Theories*. Cambridge, MA: MIT Press.
Gopnik, A. and Wellman, H. (1994) 'The Theory-Theory', in L. Hirschfeld and S. Gelman (eds) *Mapping the Mind: Domain Specificity in Cognition and Culture*, pp. 257–93. New York: Cambridge University Press.
Greene, G. (1993 [1938]) *Brighton Rock*. New York: Everyman's Library.
Harris, P.L. (1992) 'From Simulation to Folk Psychology: The Case for Development', *Mind and Language* 7(1–2): 120–44.
Heath, C. and Heath, D. (2007) *Made to Stick: Why Some Ideas Survive and Others Die*. New York: Random House.
Herman, D. (2006) 'Genette Meets Vygotsky: Narrative Embedding and Distributed Intelligence', *Language and Literature* 15(4): 357–80.
Kelley, C.M. and Jacoby, L.L. (1996) 'Adult Egocentrism: Subjective Experience Versus Analytic Bases for Judgment', *Journal of Memory and Language* 35(2): 157–75.
Keysar, B. (1994) 'The Illusory Transparency of Intention: Linguistic Perspective Taking in Text', *Cognitive Psychology* 26(2): 165–208.
Keysar, B. and Henley, A.S. (2002) 'Speakers' Overestimation of Their Effectiveness', *Psychological Science* 13(3): 207–12.
Keysar, B., Lin, S. and Barr, D.J. (2003) 'Limits on Theory of Mind Use in Adults', *Cognition* 89(1): 25–41.
Lea, R.B., Mason, R.A., Albrecht, J.E., Birch, S.L. and Myers, J.L. (1998) 'Who Knows What About Whom: What Role Does Common Ground Play in Accessing Distant Information?' *Journal of Memory and Language* 39(1): 70–84.
Leslie, A. and Frith, U. (1988) 'Autistic Children's Understanding of Seeing, Knowing and Believing', *British Journal of Developmental Psychology* 6(4): 315–24.
Leslie, A.M. and Polizzi, P. (1998) 'Inhibitory Processing in the False Belief Task: Two Conjectures', *Developmental Science* 1(2): 247–54.

Levinson, S.C. (1983) *Pragmatics*. Cambridge: Cambridge University Press.
MacWhinney, B. (1977) 'Starting Points', *Language* 53(1): 152–68.
MacWhinney, B. (2005) 'The Emergence of Grammar from Perspective', in D. Pecher and R.A. Zwaan (eds) *The Grounding of Cognition: The Role of Perception and Action in Memory*, pp. 198–223. Cambridge: Cambridge University Press.
McKee, R. (1997) *Story: Substance, Structure, Style, and the Principles of Screenwriting*. Los Angeles: HarperEntertainment.
Morton, A. (1980) *Frames of Mind: Constraints on the Common-Sense Conception of the Mental*. Oxford: Clarendon Press.
Nabokov, V. (1962) *Pale Fire*. New York: Putnam.
Narayanan, S. (1997) 'KARMA: Knowledge-Based Active Representations for Metaphor and Aspect', unpublished PhD dissertation, University of California, Berkeley.
Newton, E.L. (1990) 'The Rocky Road from Actions to Intentions', unpublished PhD dissertation. Stanford University.
Nünning, A. (1999) 'Unreliable, Compared to What? Towards a Cognitive Theory of Unreliable: Prolegomena and Hypotheses', in W. Grünzweig and A. Solbach (eds) *Transcending Boundaries: Narratology in Context*, pp. 53–73. Tübingen: Gunter Narr Verlag.
Pronin, E., Lin, D.Y. and Ross, L. (2002) 'The Bias Blind Spot: Perceptions of Bias in Self Versus Others', *Personality and Social Psychology Bulletin* 28(3): 369–81.
Sayers, D. (1946 [1929]) 'The Omnibus of Crime', in H. Haycraft (ed.) *The Art of the Mystery Story: A Collection of Critical Essays*, pp. 71–109. New York: Biblo and Tannen.
Sheehan, P. (2002) *Modernism, Narrative and Humanism*. Cambridge: Cambridge University Press.
Sternberg, M. (1978) *Expositional Modes and Temporal Ordering in Fiction*. Baltimore, MD: Johns Hopkins University Press.
Sternberg, M. (2003) 'Universals of Narrative and Their Cognitivist Fortunes (I)', *Poetics Today* 24(2): 297–395.
Thompson, J. (1964) *Pop. 1280*. Greenwich, CT: Fawcett Gold Medal.
Werth, P. (1999) *Text Worlds: Representing Conceptual Space in Discourse*. London: Longman.
Wimmer, H. and Perner, J. (1983) 'Beliefs About Beliefs: Representing and Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception', *Cognition* 13(1): 103–28.
Wright, W.H. (1946 [1927]) 'The Great Detective Stories', in H. Haycraft (ed.) *The Art of the Mystery Story: A Collection of Critical Essays*, pp. 33–70. New York: Biblo and Tannen.
Yacobi, T. (1981) 'Fictional Reliability as a Communicative Problem', *Poetics Today* 2(2): 113–26.
Zunshine, L. (2003) 'Theory of Mind and Experimental Representations of Fictional Consciousness', *Narrative* 11(3): 270–91.
Zunshine, L. (2006) *Why We Read Fiction: Theory of Mind and the Novel*. Columbus, OH: The University of Ohio Press.

Address

Vera Tobin, Department of Cognitive Science, Case Western Reserve University, 10900 Euclid Ave., Cleveland OH 44106–7068, USA. [email: vera.tobin@case.edu]