Psychopathology, Narrative, and Cognitive Architecture

(or: why AI characters should be just as screwed-up as we are)

Ian Horswill

EECS Department and Department of Radio/Television/Film
Northwestern University
2133 Sheridan Road, Evanston IL 60208
ian@northwestern.edu

Abstract

Historically, AI research has understandably focused on those aspects of cognition that distinguish humans from other animals – in particular, our capacity for complex problem solving. However, with a few notable exceptions, narratives in popular media generally focus on those aspects of human experience that we share with other social animals: attachment, mating and child rearing, violence, group affiliation, and inter-group and inter-individual conflict. Moreover, the stories we tell often focus on the ways in which these processes break down. In this paper, I will argue that current agent architectures don't offer particularly good models of these phenomena, and discuss specific phenomena that I think it would be illuminating to understand at a computational level.

Introduction

One of the things that's fun about building AI-based characters for interactive narrative is that it gives us an opportunity to model aspects of human behavior that we wouldn't otherwise have reason to duplicate in computers. Nobody particularly wants a neurotic Roomba. But there are many neurotic humans. Moreover, a disproportionate number of story characters are neurotic or otherwise a little off. People are interested in people, and we like to tell stories about people's quirks, foibles, and strengths.

There's been a great deal of progress in the last decade on architectures for agents that are believable (Bates 1994) in the sense that they display appropriate emotions, attention, etc. (e.g. (Loyall and Bates 1991), (Mateas and Stern 2002), (Blumberg 1996), (Gratch and Marsella 2001), (Reilly 1996)). However, I would argue that most current AI architectures are in some sense still too rational. I'll discuss three syndromes that are commonly portrayed in narrative, that would be challenging to model in current architectures, but that I think would be highly informative to *try* to model.

My goal is to argue that (1) we can use interactive narrative as a playground for thinking about psychodynamics in concrete, computational terms, but that (2) it won't necessarily be easy because current architectures aren't really designed for it. This isn't to say that rationality is irrelevant, that current architectures are bad, or that they can't model these sorts of phenomena at all (most are, after all, Turing complete). However, it's more than a matter of building a few new behaviors or adding domain knowledge to a reasoning system.

This is only a position paper. I'm outlining a research program, trying to explain why I think it's interesting, and suggesting some directions to look within it. I'll also make some speculations about interesting directions to look in architecture. But I don't claim to have proven anything about anything.

Example

Consider the following story fragment:

> **Scene 1:** Bill, a high school student, looks at himself in the mirror. He adjusts his tie. He changes his shirt. He changes the part in his hair.
>
> Bill: (Sighs) "Fat."
>
> **Scene 2:** A dance at the high school gymnasium. The dance floor is filled with students dancing with various levels of confidence. Others, including Bill, stand around the perimeter, watching, looking nonchalant as best they can. Bill looks across the room at Linda, who is also on the perimeter. She looks back at him and he looks away quickly. Linda looks away.
>
> **Scene 3:** Bill alone on the sofa, somewhat disheveled, watching television, eating a pint of Haagen-Dazs.

This is a variant on the standard boy-meets-girl plot in which the protagonist and love object both have low self-esteem. We can change the sexes of Bill and Linda (including making them the same sex) and it would be equally recognizable. Or we could change their ages,

perhaps even setting it in a nursing home. The story establishes that Bill has low self-esteem because of his weight (real or perceived), shows us that Bill comforts himself by eating, and that this ultimately reinforces his low self-esteem. The problem with this story from an AI standpoint is that if Bill was an AI program, the story would be more likely to read:

Scene 1: Bill looks in the mirror.

Bill: "Fat."

Cut to: montage of Bill at the health club, Bill jogging, Bill eating salads.

Scene 2: Bill and Linda dancing at the prom.

Scene 3: Bill and Linda married and living in the suburbs with dog and 2.8 children.

Closing credits

AI-Bill might be much happier. Unfortunately, AI-Bill makes a boring character for a boy-meets-girl story.[1] Even more unfortunately, the first version of Bill is often a better model of human behavior.

Self-medication

When Bill tries to drown his sorrows in ice cream, he's self-medicating. Originally, self-medication referred to patients using drugs without a prescription. For example, it's been argued that addictive disorders are best thought of as patients attempting to self-medicate underlying affective and/or neurochemical disorders (Khantzian 1997). However, the term has come to be used to refer to any sort of behavior pursued in an attempt to regulate one's own mood. When *Sex and the City*'s (Bushnell 1998) Carrie Bradshaw buys a new pair of Manolo Blahnik strappy sandals to cheer herself up, she's self-medicating. In *Romancing the Stone* (Zemeckis 1984), we're intended to see Joan Wilder's writing and consumption of romance novels as a way of self-medicating for the loneliness she feels in an otherwise isolated life. These are examples of self-medication in service of mood elevation, however people also self-medicate to lower (or sometimes raise) overall arousal. When Rick Blaine (Humphrey Bogart) in *Casablanca* (Curtiz 1942), drinks himself unconscious after a chance encounter with his former lover Ilsa (Ingrid Bergman), he's not trying to make himself happy so much as to numb himself out. Less dramatically, alcohol is used by many people to "unwind" after work, or to reduce social anxiety in professional culture (e.g. receptions at conferences) or dating.

In extreme cases, self-medication behavior becomes compulsive, leading to addiction. Wikipedia lists dozens of 12-step recovery groups for various addictive behaviors including sex, shopping, and deliberate self-injury. Such compulsive behavior is also a frequent theme in narrative. For example, the short-lived but amazing television series *Starved* (Schaeffer 2005) built an entire comedy-drama series out of characters suffering from various forms of compulsive eating, purging, and exercise.

What's interesting about self-medication is that although it is generally caused by some outside stressor, its goal is not to alleviate the stressor, so much as to regulate one's own affective response to the stressor. If you get drunk because your spouse left you, the goal of drinking isn't to get your spouse back, but to restore some sort of emotional equilibrium. Bradley has argued that the breakdown of affect self-regulation is one of the major factors in the development of psychopathology (Bradley 2000).

Whatever the mechanisms, it's clear that people engage in behaviors designed to self-regulate affect, that these sometimes degenerate into self-destructive compulsions, and that they are commonly used in narrative both as plot devices and as ways of defining a character. People engage in hedonic behavior to elevate mood, they distract or otherwise numb themselves to reduce arousal or anxiety, and they displace fear, anger, or aggression from their original objects toward new objects, e.g. by kicking the dog or getting in a bar fight, or by sublimation, e.g. by engaging in vigorous exercise, or otherwise "letting off steam".

Although we can always add a rule to an agent's program that says `(IF NO-DATE EAT-ICE-CREAM)`, current architectures don't account well for the pervasiveness and systematicity of this general phenomenon.[2] Not everyone will eat ice cream in response to perceived rejection, but nearly everyone will respond with some sort of self-soothing behavior. What varies from individual to individual is the type, intensity, and duration of such behaviors, as well as their thresholds for activation.

Limerence

An enormous fraction of popular narrative involves some kind of love plot. These sometimes involve an external threat to a protagonist's existing relationship, but very often (especially in Hollywood narrative), they involve the formation of a new romantic relationship. New relationships typically involve more uncertainty, and therefore more opportunity for dramatic conflict. Thus a very large fraction of narrative involves somebody falling in love with someone else.

[1] This isn't to say we might not see Bill eventually diet and win Linda over, but this would be his resolution to the conflict, which would come in the third act. It wouldn't be the complete story.

[2] Consider the fact that the developed world is fighting simultaneous epidemics of obesity and anorexia.

Limerence is the term proposed by Tennov for the pattern of obsession, idealization, and fear commonly associated with "falling in love" (Tennov 1999), most importantly:

- Intrusive ideation (thoughts) about the object of desire (the limerent object[3], or LO)
- Acute longing for reciprocation by the LO
- Shyness of and fear of rejection by the LO
- Hypersensitivity to, and dependence of mood on, cues for acceptance or rejection by the LO

Although limerence often lasts for years, Tennov distinguishes between it and "mature love", which involves the development of a stable attachment and concern for the other. She also distinguishes limerence from sexual attraction, which while usually involved, is not sufficient for limerence.

One of the most paradoxical characteristics of limerence is that it is driven in large part by uncertainty. People don't become limerent toward those who indicate unambiguous interest or rejection toward them, but toward those whose behavior is ambiguous or inconsistent (Tennov 1999). Once limerence begins, perceived rejection by the beloved actually increases the amount of time spent in limerent fantasy rather than reducing it (although sustained rejection will reduce and ultimately eliminate it). It is similar to, and possibly related to, the phenomenon in operant conditioning where inconsistent reinforcement schedules generate a stronger response than consistent ones.

Limerence can reduce us to blithering idiots, often literally. One of the running gags in the television series *South Park* (Stone and Parker 1997) is that Stan Marsh, one of the protagonists, throws up every time he tries to talk to his would-be girlfriend Wendy Testaburger. Some people stammer when trying to talk to their limerent object. Others become so shy they can't even approach their beloved. And many find their minds have gone blank and can't think of anything to say, or conversely, become impulsive and blurt things they've only half thought-out. This is presumably related to the phenomena of social inhibition in which people perform worse on unfamiliar tasks when being watched by others. Social inhibition is believed to occur because social attention acts as a distractor, effectively adding a short-term memory load.[4]

Limerence is a form of obsession. Subjects experience intrusive ideation, meaning their attention repeatedly returns to the limerent object, even when they attempt to think about other topics. Intrusive ideation isn't limited to limerence, however; it's a common reaction to situations of extreme affect, including grief and guilt. It's also associated with a range of psychiatric disorders, including paranoia and other delusional disorders, post-traumatic stress disorder, stalking, and suicidal behavior. Yet it isn't even clear what it would mean for most AI systems to become obsessed in this way. Many agent architectures, such as Soar (Laird, Newell et al. 1987), Hap (Loyall and Bates 1991), and ABL (Mateas and Stern 2002), shift attention between goals and subgoals rather than between objects. Object obsession seems easier to model in systems like Blumberg's Hamsterdam (Blumberg 1996) or *The Sims* (Wright 2000), both of which are organized around competing drives and objects toward which those drives can be discharged. The interesting challenge would be to integrate such a "low-level", ethologically oriented architecture with the "higher-level" reasoning and problem-solving capabilities of traditional AI architectures.

Limerent ideation usually takes the form of elaborate fantasies culminating in the beloved's indication that s/he reciprocates the subject's love (Tennov, 1999, p. 39). These fantasies often involve some self-sacrifice on the part of the subject, even a heroic death that simultaneously earns the beloved's reciprocation while removing the risk of its future loss. Again, this is interesting for AI because current architectures don't allow agents to satisfying goals (albeit temporarily and unsatisfactorily) through fantasy. Yet this is absolutely key, not only to love plots such as in Romancing the Stone, but to actual human behavior.

The final aspect of limerence that bears mention is hypervigilance: the extreme sensitivity to cues regarding the beloved's attitudes toward the subject. Again, hypervigilance isn't limited to limerence, it is also commonly found in adult children of alcoholics, abuse survivors, PTSD patients, and paranoids. It's interesting to note that Colby's PARRY system (Colby 1973) was able to produce a relatively convincing simulation of paranoid belief without including an actual reasoning system; PARRY was essentially a modified version of Wiezenbaum's ELIZA system (Wiezenbaum 1966), with additional components that performed a surface-level analysis of the text to determine the interlocutor's level of aggression toward the program (hypervigilance) and to detect the occurrence of trigger concepts that would grab the system's attention (obsession).

Authoritarian personality

Although not nearly as common a theme as love, a great deal of late 20th century popular fiction explores the themes of fascism and authoritarianism, especially in science fiction and fantasy. For example, in J.K. Rowling's Harry Potter novels, the wizarding world is threatened by the Death Eaters, who seek to impose a new authoritarian order led by Lord Voldemort. The Death Eaters value racial purity, class, and the pursuit of pure power (Rowling 2000) Among the students of Hogwarts School, the followers of Voldemort are typified by the

[3] This is "object" in the sense of object relations theory, not in the sense of (mere) physical object.

[4] This is suggested by the fact that performance doesn't degrade on well-rehearsed tasks, and can actually be improved by an audience, a phenomenon known as social facilitation.

character of Draco Malfoy, an aristocratic student with strong racist and classist prejudices. Malfoy is portrayed as being a bully to those below him in the social hierarchy and fawning toward those above him. Suppose we wanted to modify the Harry Potter games so that Malfoy was an autonomous NPC, rather than being almost entirely pre-scripted. What, exactly, would we be simulating?

There has been considerable research into the personality traits associated with following authoritarian leaders. The California Fascism Scale (Adorno, Frenkel-Brunswik et al. 1982) introduced a set of personality traits thought to underlie authoritarian personality, in the sense of one who is attracted to following authoritarian leaders. Later, Altemeyer (Altemeyer 1996) analyzed the Berkeley group's work and found it could be explained almost entirely by the conjunction three of the personality traits: conformity, submission to authority, and aggression on behalf of authority. These three traits covary surprisingly strongly, with alpha values in the 0.8-0.98 range, making it one of the stronger results in social psychology (*ibid*, pp. 18-19). Moreover, they predict a number of other attitudes such as nationalism, ethnocentrism (racism, classism, etc.), intolerance toward homosexuals, and political affiliation[5] (pp. 21-31). Authoritarianism also predicts certain aspects of behavior. For example, when asked to choose punishments for others' crimes, high authoritarians will in general choose more severe punishments than low authoritarians, and report greater pleasure in administering the punishment. However, their assignment of punishment will depend on the identity of the perpetrator; an accountant who started a fight with a "hippie panhandler" will be given less of a punishment than if the subject is told the hippie started the fight with the accountant (p. 23). Interestingly, while authoritarianism does not correlate with IQ scores, it does correlate with failure to recognize faulty reasoning in others (pp. 94-95) and compartmentalization, the tendency to unknowingly hold contradictory beliefs in different circumstances (pp. 95-101).

The Harry Potter books portray all these traits and behaviors in both Draco Malfoy and Dolores Umbridge (another authoritarian portrayed as evil). The only exceptions are compartmentalization, failure to recognize faulty reasoning, and intolerance toward homosexuals, which are never addressed in the books. Modeling these traits computationally requires build reasoning systems in which (1) reasoning processes depend on the social status and affiliation of those being reasoned about, yet (2) the system itself is unaware of such dependencies.

[5] Interestingly, while high authoritarians tend to be Republican in the US and Conservative in Canada, they tend to be pro-communist in Russia.

Men are dogs (no, really, I mean it; women too)

Authoritarian personality is interesting because it provides an extreme example of common behavioral patterns having to do with group affiliation, and the ways in which those processes interact with "higher level" processes such as reasoning. I would argue that for many purposes, humans basically act like dogs with large forebrains: we break up into packs, the packs tend to have status hierarchies, and we tend to defer to the top dog in the pack. One interesting difference from other mammals, however, is that humans are often members of many different packs at once (one's family, department, church choir, etc.). Moreover, our membership and status within a pack are often based in part on behavior, that is, on conformance to group norms.

Again, consider the case of the Harry Potter novels. As with most high-school novels, groups and belonging are major themes. Readers of the Harry Potter novels have to keep track of at least two dozen different groups of 6 different types:

- Quiddich teams and their supports (Quiddich being a kind of aerial soccer/football)
- Magical schools (Hogwarts, Beauxbatons, Durmstrang)
- Houses within Hogwarts (Gryffindor, Hufflepuff, Ravenclaw, and Slytherin)
- Paramilitary groups (for want of a better term; The Death Eaters, The Order of the Phoenix, the "DA", the Inquisitorial Squad)
- Quasi-racial categories (wizards, pure blood wizards, blood traitors, muggles, and squibs)
- Social classes (working class, professional, middle class, upper class)

along with their membership and leadership, and conflicts, just to understand the social lives of the main characters

For example, the character of Ron Weasley is a middle class, pure blood wizard, a supporter of the Bulgarian quiddich team (at least at first), a student of Hogwarts, in which he is prefect of House Gryffindor, a member of the DA, and a sort of auxiliary member of the Order of the Phoenix. Most of these group types (teams, school, houses, the paramilitaries, etc.) are in explicit competition or conflict; each group competes with the other groups of that type. To understand Ron's interactions with Malfoy, the reader has to understand that they're involved in at least 5 different pair-wise conflicts: Gryffindor vs. Slytherin, Order of the Phoenix vs. Death Eaters, DA vs. the Inquisitorial Squad, blood traitor (in Malfoy's eyes) vs. pure blood, and middle/professional class vs. upper/aristocratic class. If you haven't read the Harry Potter novels, and so find this hard to follow, I apologize. However, it only strengthens my point.

A central source of both conflict and humor in the books is the changing system of allegiances of the different

characters, and the ways they affect the characters' judgment without their realizing it. Ron Weasley idolizes Bulgarian quiddich player Viktor Krum until (1) Viktor's school is placed in competition with Ron's and (2) Viktor competes with Ron for Hermione Granger's affections, after which Ron can see no redeeming qualities in whatsoever in Viktor. It is an important part of Ron's character that he is in some sense unaware of his attitudes having changed, much less of the reasons for them having changed.

Psychopathology, narrative, and cognitive architecture

I've tried to argue here that there are a number of common phenomena of human behavior that would be viewed as failures in a robot, but that are recurring themes in narrative. Self-medication is an example of a behavior involved in affect regulation, which most current architectures don't include as a possible form of goal or motivation. Limerence is an example of both obsession, which is easier to model in more ethologically oriented systems like Hamsterdam and *The Sims*, and of fantasy, which would appear to require quite an elaborate cognitive system capable of simulation and planning, although possibly different from current planners. Authoritarian personality provides a (sometimes extreme) example of a set of personality traits that appear not infrequently in narrative, seem to have psychological reality, and involve a complicated dependence of reasoning processes on social factors such as group affiliation and rank, of which the subject is unaware.

Again, none of this is to say that people are fundamentally irrational or that logical reasoning has no place in believable agents. Rather, we need to develop cognitive systems whose "failure" modes are closer to those of humans, since those are often the situations about which we choose to tell our stories. Nor is this to say that phenomena such as limerence, affect regulation, or group loyalty are failure modes *per se*. But we can learn a great deal by looking at the cases where these do lead to pathology, just as we can learn about vision by studying optical illusions.

Unfortunately, this is a position paper arguing for a research program, not a report of real results. I don't claim to have a better architecture. But I would argue that human cognitive architecture must have innate support for certain forms of sociality such as group affiliation and social status, if only because breakdowns of it such as paranoia can be induced by hardware-level changes such as certain forms of brain damage or the use of amphetamines. Four of the five subforms of delusional disorder listed in the APA's Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric Association 2000) specifically relate to beliefs about social status or interactions. Moreover, serious social deprivation, such as occurs in solitary confinement, produces severe psychotic symptoms including hallucination and paranoia in surprisingly short periods of time, and results in long-term personality changes (Grassian 2006).[6] And, again, social status and group membership unconsciously affect reasoning processes.

To me, this suggests a "dogs with large forebrains" model in which the neural substrates of social behavior, which we've inherited from other social mammals, are still active in humans, but have been repurposed as motivational systems and as inputs to reasoning rather than as direct generators of behavior. Whether this is the right model or not, I think we can learn a lot, and have a whole lot of fun, by trying to make artificial characters that share our neuroses.

Acknowledgements

I'd like to thank Lauren Berlant, Doug Church, Robin Hunicke, Christine Lisetti, Andrew Ortony, Robert Zubek, and the reviewers for their comments, criticisms, and encouragement. I hope my revisions have been as illuminating as their comments.

References

Adorno, T. W., E. Frenkel-Brunswik, et al. (1982). *The Authoritarian Personality*. New York, Norton.

Altemeyer, B. (1996). *The Authoritarian Specter*. Cambridge, Mass., Harvard University Press.

American Psychiatric Association. and American Psychiatric Association. Task Force on DSM-IV. (2000). *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV-TR*. Washington, DC, American Psychiatric Association.

Bates, J. (1994). "The Role of Emotion in Believable Agents." *Communications of the ACM* **37**(7): 122-125.

Blumberg, B. (1996). Old Tricks, New Dogs: Ethology and Interactive Creatures. Ph.D. thesis, MIT Media Lab. Cambridge, Massachusetts Institute of Technology.

Bradley, S. (2000). *Affect Regulation and the Development of Psychopathology*. New York, Guildford Press.

Bushnell, C. (1998). *Sex and the City*. C. Bushell, D. Starr and M. P. King, HBO Films.

[6] Note, however, that it is hard here to separate the effects of social deprivation from the sensory deprivation that also occurs in solitary confinement.

Colby, K. M. (1973). Simulation of Belief Systems. *Computer Models of Thought and Language*. R. C. Schank and K. M. Colby. San Francisco, W.H. Freeman and Company**:** 251-286.

Curtiz, M. (1942). *Casablanca*. USA, Warner Brothers.

Grassian, S. (2006). "Psychiatric Effects of Solitary Confinement." *Journal of Law & Policy* **22**: 325-383.

Gratch, J. and S. Marsella (2001). Tears and Fears: Modelling Emotions and Emotional Behavior in Synthetic Agents. *Proceedings of the 5th International Conference on Autonomous Agents*. Montreal, ACM Press.

Khantzian, E. J. (1997). "The Self-Medication Hypothesis of Substance Use and Disorders: a Reconsideration and Recent Applications." *Harvard Review of Psychiatry* **4(**5): 231-244.

Laird, J. E., A. Newell, et al. (1987). "Soar: An Architecture for General Intelligence." *Artificial Intelligence* **33**: 1-65.

Loyall, B. A. and J. Bates (1991). *HAP: A Reactive, Adaptive Architecture for Agents*. Pittsburgh, Carnegie Mellon University School of Computer Science.

Mateas, M. and A. Stern (2002). "A Behavior Language for Story-Based Agents." *IEEE Intelligent Systems* **17**(4): 39-47.

Reilly, S. N. (1996). *Believable Social and Emotional Agents*. Ph.D. thesis, Computer Science Department. Carnegie Mellon University. Pittsburgh, PA,

Rowling, J. K. (2000). *Harry Potter and the Goblet of Fire*. New York, Arthur A. Levine Books.

Schaeffer, E. (2005). *Starved*, FX Network.

Stone, M. and T. Parker (1997). *South Park*, Paramount/Comedy Central.

Tennov, D. (1999). *Love and Limerence : The Experience of Being in Love*. Lanham, MD, Scarborough House.

Wiezenbaum, J. (1966). "ELIZA." *Communications of the ACM* **9**: 36-45.

Wright, W. (2000). *The Sims*, MAXIS/Electronic Arts.

Zemeckis, R. (1984). *Romancing the Stone*. USA, 20th Century Fox.