A.I. and Emotional Robots: Collaborative Fiction in Science and Film

Peter Asaro

Roboexotica November, 2005
Museumsquartier Vienna, Austria

When I was first asked to give a paper on the topic of "Collaborative Fiction" I knew immediately that I would want to write about the film *AI Artificial Intelligence*. It was, after all, one of the more complicated stories of authorship in filmmaking. It was also a film that came out at a time when I was making my own independent documentary film on nearly the same subject, *Love Machine*. In this paper I want to discuss some of the complexities of the collaborative fiction of the "AI" film, and the kinds of collaborative fictions that we find in the scientific practice of AI, as well as the complexities of the collaboration that is our social reality and what it might mean for a robot to share in it.

AI, both the film and the scientific discipline, are highly collaborative enterprises. Of course, all large Hollywood films incorporate the work of many individuals, and are thus essentially collaborative. Similarly, scientific projects are generally pursued by a collaborative effort within the laboratory, and usually these depend heavily on the work of other labs, if not direct cooperation. But there is a sense in which this film is more collaborative than other projects. Most obviously, the film was a project pursued by the director Stanley Kubrick for many years, though it was Steven Spielberg who ended up actually directing the film. This shift in directors was motivated primarily by Kubrick's death, and has been the subject of quite a bit of controversy and discussion.

The story goes like this: Kubrick met the science fiction author Brian Aldiss in 1973. An admirer of the *2001: A Space Odyssey* film, Aldiss presented Kubrick with a book of his short stories. Among those short stories was one in particular that fascinated Kubrick. Entitled "Supertoys Last All Summer Long" it was about an artificial boy who has a robotic toy teddy bear but does not realize that he too is robotic. But it wasn't until ten years later, when Kubrick saw Spielberg's film *E.T. the Extra-Terrestrial*, that he was inspired to make another science fiction film and eventually bought the film rights to "Supertoys" in 1983.

The process of turning the short story into a screenplay was itself a collaboration between many authors. Of course, Kubrick had strong influence on the rewriting, which was first

undertaken by Aldiss, and at times also by Kubrick's co-author for *2001*, Arthur C. Clarke, as well as Bob Shaw, and novelist Sara Maitland. A completed version of the screenplay based on these efforts was finally produced for Kubrick by sci-fi author Ian Watson. But "Supertoys" wasn't the only source for the story, as it also draws heavily upon the fairy tale story of "Pinocchio," and indeed Kubrick always referred to it as his "Pinocchio" story. So we ought to add to our list of authors the name Carlo Collodi, which is the pen-name of the 19th century Italian writer Carlo Lorenzini. Our list of authors is already up to eight, and the screenplay has not yet been turned into a shooting script. Upon Kubrick's death, all the pre-production materials were transferred to Spielberg, who then reworked and completed the script. Ultimately, in the opening credits of the film, Spielberg and Ian Watson share in the writing credit, and it was Spielberg's first writing credit since *Close Encounters*. And so we have already the text of a story with no less than nine authors, and some sense of how the work of so many people is often reified in credit for just a few. I can further assure you that the development of the visual story involved numerous cinematographers, production designers, graphic and scenic artists, costume, make-up and special effects masters working with both digital and robotics techniques, and the powerful cinematic vision of two of the great filmmakers of the 20th century, Kubrick and Spielberg.

There was initially some public surprise at Spielberg's involvement, and a sense that he had hijacked the film. It turns out that Kubrick and Spielberg had discussed the film at length and numerous times, beginning in the early 1980s. Ultimately Spielberg decided to finish the film which his own work had inspired in Kubrick, as a way of paying tribute to his friend and fellow artist.

If all that isn't enough collaboration for you, after the film's release on DVD, it became subject to a phenomenon called the "fan re-edit" It seems a filmmaker in Sacramento, CA, named DJ Hupp, decided that he could edit the film to be closer to Kubrick's vision, based on some news articles he read, and thus re-edited the theatrical release of the film, eliminating some 20 minutes from the running time, but adding no new shots. Mostly he eliminated elements he thought were silly, in an effort to give the film a darker feel.

The fact that a film is a collaborative project is not very remarkable in itself, but what about the content of the story? How is the AI of the protagonist boy-robot David itself construed,

and how is social collaboration implicated in this character study? The story of the film for those who haven't seen it (though I recommend it and apologize if I spoil the ending for you) goes roughly like this: In the future, human reproduction is tightly regulated by the government, and robots are used extensively for labor. Our classic mad-scientist type, Professor Hobby of Cybertronics Manufacturing, envisions building a child robot to fulfill the human need for children for those parents who are not allowed to have an organic baby. A married couple, Monica and Henry, have an organic son, but he is in a medically-induced coma, awaiting a cure for an untreatable life-threatening illness. In lieu of their real son, they accept David, a prototype of the child robot. David has a special circuit designed to love its parents through an irreversible imprinting process. Monica chooses to become imprinted as David's mother, though Henry never makes the commitment and does not get imprinted. Eventually their "real" son Martin is awoken from his coma, is cured, and returns home to harass, trick, and taunt the robot-boy until his parents are convinced that the robot-boy David is dangerous and must be returned to Cybertronics for destruction. But Monica cannot bare to destroy the robot-boy whom she's developed an affection for, and so instead abandons him in the woods.

David, abandoned in the woods with his robotic teddy bear companion, sets out on a fairy tale quest. Having been read the story of Pinocchio at bedtime, he comes to believe that if he finds the Blue Fairy, she can turn him into a real boy the way Pinocchio was turned into a real boy, and then his mother will love him. I'll return to David's story in a moment.

The real scientific practice of AI is also a collaborative fiction, a conspiracy of numerous actors. Different labs pursue different lines of research. There are fairy tale quests to build new technologies that fulfill needs that do no yet exist. There is a certain suspension of disbelief, and willingness to reconceive fundamental terms and meanings, that is required for science to progress. But what motivates this research? Indeed, why do scientists working today seek to build robots with emotions? With social skills? And more importantly how do they go about doing this, and what does it tell us about how we understand ourselves as emotional and social beings? These are deep and philosophical areas, and not ones that many engineers spend too much time worrying about.

The most obvious fictions of AI research in the past were the exaggerated claims of what machines would be able to do in just a "few short years." Rarely have these speculative

achievements been met in the number of years predicted. The problems turn out to be more complex, the techniques more limited, and, despite the exponential growth in computing power, the data processing requirements still too great.

As for the fairy tales that motivate the building of humanoid robots, we should ask "Why should we need machines to care about us?" Don't we have enough trouble with human emotions, and emotionless machines, already? Or do we really have so much trouble loving one another that we need to build some machine to do this for us?

Emotions are by their very nature a sort of collaborative fiction–a socially shared reality. Philosophers have discussed this since at least Ludwig Wittgenstein's comparison of having a pain to having a "beatle in a box" that no one else could see but you. The pain sensation is a private experience, but the expression of pain in a scream is public, shared and collaborative. Like language, these emotions have a life of their own in the public sphere, and become meaningless or inconceivable when reduced to a solipsistic private sphere. This raises the philosophical question "If a boy cries in the woods, and there is no one around to empathize, does he have emotions?"

When Monica abandons David in the woods, she is obviously emotionally upset, and does respond to his cries–not with cold indifference, but rather with an emotional unease over her own regretful actions. And while the narrative of the film requires us to see David's love as unrequited and unfulfilled, deferred until his mother takes him back, it is in this scene that his emotions are most real, most public and shared, and, most importantly, most clearly acknowledged and reciprocated by his mother. This is despite the fact that they are emotions of fear, suffering and regret, rather than happiness and joy. It is here that, despite the cruelty enacted upon him, the other characters in the film most empathize with him, especially his mother.

In the rest of the film, David does much to undermine our confidence that he truly is an emotional being, failing to collaborate with those around him or to create an emotionally rich world. Spielberg would have us believe that David's emotions are there by "special creation" via the special Cybertronics chip. This is what guarantees his emotion, and not his social interactions and development, much less his participation in a community that recognizes his emotions as his own.

Early on, David insists that his best friend, a robotic teddy bear, is not the "child" of Monica the way that he and his brother Martin are. Yet Martin torments him, eventually getting him kicked out of the house, and the teddy bear remains a helpful and dedicated friend throughout the story. The same is true of David's other companions and would-be friends. Even while David develops a strong dependence on Gigolo Joe, the sex robot, he hardly acknowledges the friendship as valuable or meaningful– they meet when David clings to him instinctually out of fear when David is about to be publically executed in an orgy of robot destruction called a "Flesh Fair." The Flesh Fair destroys old robots in a celebration of "Life." David is ultimately spared by the crowd, who again are compelled to empathize with David on a superficial level–a cute boylike robot surely does not deserve to die a terrible death. Joe is spared too because David won't let go of him, and ends up assisting in David's quest, but it is Joe who demonstrates social and emotional skills of nurturing and teaching (despite lacking an "emotion chip"), while David pursues his singular goal.

David's salvation in the "Flesh Fair" sequence also raises the question of the public sphere in determining membership in a community. "Mecha don't plead for their lives" shouts a member of the outraged crowd to the MC who is about to dump a vat of acid on David. The other robots are clearly struggling and running for their lives, in countless acts of self-preservation, yet they do not explicitly beg or plead for their lives. So what is the difference? Is it one of reasoned self-awareness or one of emotional intelligence? Clearly a plea for life seeks an emotional response from the other agent who has the power over life and death in the situation. Whether it targets compassion and mercy, or threatens the potential suffering of negative consequences, a bad conscience or a judgement of low moral esteem from the larger public, it fundamentally seeks to move *emotionally* the agent in power.

This is perhaps one of the most powerful images of collaborative fiction in the film. The reality of David's emotional life is here determined by the crowd. Pain alone is not sufficient to prohibit the violence against the robots, nor is being animate or "alive." The other robots feel pain, as do animals kept for pets and food (surely dogs and other pets seek the love of their owners too), but these creatures can be legitimately killed for utility or entertainment without the moral weight attached to killing a human. What makes the robot deserving of human rights, and lends it this moral weight? Is it a property of the robot, recognized and confirmed by the crowd

(Hobby's innovative chip)? Is it merely the whim of the crowd? Or is there some more complicated reality at work?

To say that we can identify this property, define it and seek to realize it technologically is in fact the collaborative fiction that motivates current research in emotional robots. It presents all the necessary elements: an analytic challenge to which one can employ philosophical thought experiments, develop mathematical models of psychological processes, and engineer robots to realize these models. And this is pretty much how research now proceeds, at least in outline.

However, we could also say that the whole notion of determining membership in the club of those with "human rights" is a collaborative fiction motivated by enlightenment ideals of human subjectivity. Humans kill other humans with remarkable frequency. There have been numerous human cultures that have used human sacrifice in displays of religious piety and political power, as well as displays of justice and in various modes of entertainment. The powerful aspects of human sacrifice derive from the terror of such displays, the empathetic reflexes and sense that the cause to which the life is sacrificed must be greater than the magnitude of the individual human lives lost in the display. These displays function within their cultures in ways such that the taking of human life was representative of the power of the agents taking that life–be they priests, kings, judges or gladiators–as well as the importance of their respective gods, laws, nations, ideals and bodies. It is only the modern liberal conception of humanity that seeks to set absolute moral rights upon the human subject, and which is thus confronted by the dilemma of what constitutes the human. This dilemma is now being played out in sci-fi morality plays and ethical debates over technological innovations.

But as Bruno Latour has said, "we have never been modern." There may not be such clean lines between the "matters of fact" and "matters of concern" on the subject of robotic emotions, any more than there are on the subject of "human rights." Indeed, it seems that much of the discourse draws the spurious distinction between artificial and real. But these terms are not opposed. Rather, artificial is distinct from natural, both are "real." The distinctions is not about the performance of the present, but a matter of the origin of things. And if we are to believe Latour, that there is no ontological distinction between matters of fact and matters of concern and one is just the frozen version of the its more fluid counterpart, then the emotional world of robots can also be seen as moving between the tangible emotional phenomenology of

the robot, and the fluid social categories we ascribe to them. These are a matter of concern only when we argue about them, but it seems likely that when working emotional robots really do appear, many people will accept them naturally, without question, as do some characters around David, like the little girl who notices him at the Flesh Fair. In the end, we realize the story is not about David, but about those around him. Emotion is not a chip in our brains, but a collaborative fiction, a social reality.

The film would have us believe that the most fundamental and essential form of love is mother-child love. Yet elements of the film also works against this, at the same time revealing something of our culture of parenthood. For Monica, there is a conflict in loving David as her own son because he is not the blood-of-her-blood. David is adopted. Moreover, when her real son Martin eventually returns, he is loved freely and without question, even as David longs for that same love. Still, Monica insists on treating both boys fairly, but out of a sense of justice rather than love, though she eventually shows signs of increasing fondness for David. It is Henry, the ostensive adoptive father who brought David home, that truly fails to love, though David is unconcerned with obtaining his love because Henry did not imprint himself on David. David's world is that of a modern dysfunctional nuclear family, complete with unloving step-father.

The failure of David to earn the love of Henry is the problem of the adoptive father who does not accept the new son as his own. The real father is Professor Hobby, who we later learn created David in the image of his own dead organic child. It is he who takes pride in the creation, who projects his own hopes and dreams onto David, as a father would. David's real problem is not his own emotions, but the failure of the family unit–to bond, to trust, to love one another. And this causes him to act out, in different ways perhaps than a real boy, but no less purposefully. His real father-creator, Prof. Hobby, loves David unconditionally, and loves him for precisely what he is, an artificial boy, rather than what he pretends to be through imitation, as Monica tried to love him.

To what extent is Prof. Hobby like AI researchers today? The film opens with an inspirational speech by Prof. Hobby about how he plans to revolutionize robotics, by introducing true emotions. The speech is a rallying cry to his engineering and marketing staff. It is a business plan that is enabled by this new technology, and thus demands it. This is the stuff that

drives even theoretical work today. It identifies a human need, and defines a potential market for the new technology. Is this really so different than the billion-dollar multi-institution research project in Japan to develop robots to nurse the elderly of the baby-boom generation? In the current society, the elderly need care that is no longer provided by extended families, and there is a predicted increase in this demand. There are presumably other ways to fill this need, so what justifies the need to pursue a difficult, complicated and expensive technological solution? I submit that it is the collaborative fiction of science-fiction fantasies, mixed with the occasional gee-whiz technology demonstrations, and inspirational speeches of visionaries and industrial leaders; it is the economic interests of companies, the political interests of funding governmental agencies, the cultural terrain of society, and the aspirations of scientists and engineers. What else could move culture in these new directions?

Of course, there is a substantial number of researchers who would deny that there is anything particularly profound or sensational in their attempts to make computer and robotic interfaces more friendly and responsive to natural forms of human communication. I would submit that they too are engaged in a certain collaborative fiction of denying the potential implications of their work–dues perhaps to their own latent fears of the uncertainty that technological change brings to society.

We should also ask ourselves why is mother-child love considered to be the ultimate or fundamental form of emotion? It might appear as the first form of love from which others grow, the "primordial" love. But there are other forms and it is their remarkable variations and combinations that make up the human experience. Indeed, as the head of the sex-robot division notes during Hobby's rallying speech, "we already have pleasure mechas." Why is physical love not considered real love? Or why is it a lesser form? Again we seen in Gigolo Joe's seduction of women that he knows just what to say to enact a collaborative romantic fiction, from setting the mood with music to the powerful uses of flattery. But we are meant to believe that seduction and lust are simply poor imitations of romantic love, easily imitated by sleaze-balls and low-tech robots.

In another sense, David's failure is not the failure to earn the love of his mother, but his own failure to love anyone else around him. His friends and companions are regarded with

indifference. David represents the imperfection of his creation–more precisely the imperfection of his creators. The singularity of the mother-child love is unable to develop into friendship, into caring for those around him, or into mature sexual love. His love is frozen and fixated in a sort of infinite Freudian loop. Were he human, he would surely have grown up to live with his mother, never asking anyone out on a date or exploring the world. But this is precisely what Prof. Hobby's chip made him to be. He was designed to be emotionally limited in order to meet the perceived desires of the consumers of child-robots.

Pinocchio wanted to be a real boy, but his adventures rested in his doing remarkably original and precocious, if silly or naughty, things. That is, his behavior was all-too-human, but his body was wooden. David is the reverse, his body is extremely humanoid, but it is his odd behavior that sets him apart as not having human emotions. Which brings up another bit of interesting collaborative fiction. In the original publication of Pinocchio as a magazine serial, Pinocchio was killed in chapter 15, hung by assassins for not doing what they told him to do. Though unlike David in the Flesh Fair, this killing was meant to weigh as a moral burden on the thieves who kill Pinocchio, and thus his execution testifies to his humanity, if only in the reader's reflection on the immorality of his murderers. Upon his editor's urging, Collodi rewrote the 15th chapter, and added chapters 16-36 to the story, in which Pinocchio seeks out the Blue Fairy in order to become a real boy. In a sense, these chapters are tacked on to give a more agreeable ending, to turn a poignant tragedy into something hopeful. It parallels the oddly tacked-on ending of the film, which many reviewers of the film have criticized.

Upon learning his true robotic nature in the midst of a warehouse of copies of himself, and thereby learning that he is not unique (a clearly human recognition and fear, as opposed to the more optimistic notion that humans are all unique and special) David attempts suicide, but is saved by Joe. He then finds his Blue Fairy in a Coney Island fun-ride submerged below the ocean, and repeatedly begs her to make him real until his batteries die. He is then resurrected 2000 years later by some super-race of robots, for whom he was a sort of Adam, cast out of the garden. They want to honor his wish and satisfy his desire, and thus resurrect his mother too, if only for one day, so that she can love him once more. And this satisfies him. This ending surely fails as a morality tale, since it tells us that singular determined unending devotion to a cause is the only hope of success, even when it is clear that the odds are ridiculous, or the desired goal

becomes less desirable through a developing maturity. For a living creature, a loving creature, it must be possible for one's goals, and one's emotions, to develop, to change, to take different objects. Humans often become angry with those they love, they must learn to accommodate difficulties, and must learn to love in different ways. David achieves none of these capabilities.

But what is the further cultural significance of this "special creation" myth that David embodies? Why is it that only the "pure" and "essential" human form is deserving of rights and respect? Animal rights advocates have long sought to challenge such notions. But their opponents are quick to point out that surely we don't want to give moral respect to mere objects–trees and rocks, cars and computers–do we? That would be absurd! To this one might respond that, well, perhaps a horse is not as important as a person, but it is still deserving of some respect, compassion and empathy, though perhaps not as much as a pigeon in the park. Without trying to argue that there is some clear-cut hierarchy, it seems reasonable to suppose that it might be useful to consider a continuum, instead of a binary opposition, of moral values such that inanimate and animate objects take up places deserving of some respect, and in large numbers perhaps even deserving of rights. I believe environmental ethicists have proposed something like this in the concept of the "natural contract" in which a forest, for instance, is deserving of some moral esteem, or at least some serious and respectful consideration before being clear-cut. Indeed, there is a subtler moral motif running through the film, set in a world drowned by the rising waters of global warming. And that motif speaks to the social and environmental costs of preserving and reinforcing the myth of special creation–the demonstration of human control over the non-human through continual and conspicuous exploitation.

In several ways the arguments about what is essentially alive, emotional or human are caught up in the modern liberal conception of what it means to be a moral and legal agent, the bearer of intrinsic human rights. We can thus see the "problem" of whether robots deserve rights not as a litmus test for our philosophical definitions, but as an opportunity to reconceive moral and legal subjectivity. I believe that this promises to be a challenging enterprise, and one which scientists and technologists will in all likelihood shy away from.

The common thread of many science fiction robot tales is that the children are our future, and it is the robots made in humanity's own image who shall inherit the world of the future. It is thus up to us, the last remaining human generations as it were, to engineer the moral fabric of

that world, to design robots not merely in our own image, but in the image of what we would like to be. Needless to say, Spielberg's moral vision tends to be powerful, though rather limited in its dynamics and scope. Kubrick's is deeper, though more disquieting and generally darker. Of course, the future will need both the bright and dark, and will inevitably be the collaborative fiction of the whole of culture.