A Proposal for the Humanoid Agent-builders League (HAL)

Joanna Bryson
Division of Informatics; University of Edinburgh; UK
joannab@cogsci.ed.ac.uk

Abstract

The following is a proposal for the Humanoid Agent-builders League — a professional organisation for people responsible for creating artificial people. Although the league would have all the advantages and enjoyment of any professional organisation, its main function would be to create and maintain an ethical standard for the field, both with respect to its consumers and its product.

1 Proposal

1.1 Introduction

This proposal is intended to address the unique ethical issues associated with the intentional creation of human-like artificial agents. In our society, there is significant pressure from economic reward to create agents that exploit human social drives. We consider such exploitation potentially unethical in that it can be damaging to human lives, human society and other vital concerns. This is because the exploitation misappropriates evolved inclinations to devotion, directing it towards objects that do not actually require extensive resources. Since many of the resources of individuals in society are relatively fixed (particularly time), such misappropriation costs society as a whole, as other worthwhile and needy recipients will go wanting.

We seek to address this problem by creating a professional league for the developers of humanoid agents. This solution is inspired by the Magicians' Unions, which take advantage of the prestige and expertise endowed by formal membership to further basic ethical standards for practitioners of their craft. While not entirely preventing the existence of charlatans such as "faith healers" from exploiting the public, the Magicians' Unions perform a considerable service. They both educate their own members as to their ethical responsibility, and serves as a platform for educating the public about the deceptive power of professional magic. We hope the Humanoid Agent-builders League (HAL) could similarly server both as an entertaining and informative club, and as a source of social good.

1.2 A Code of Ethics

We propose a three element ethical code of practice for the Humanoid Agent-builders League. This code takes the form of a series of promises made to our consumers.

1. **Honesty** (the Right to Knowledge): No consumer should be falsely persuaded that the requirements of an artificial agent are in any way equal in importance to the needs and desires of humans or animals. Consumers should not be coerced to spend time, money or energy for the benefit of the artificial agent, but only for their own enjoyment. It should be made clear on all products that the apparent joy or suffering of the agent are devices manufactured by a human programmer for the advantage of the consumer. On adult products, this can take the place of a standard disclaimer along the lines of the Copy-Left agreement used by the Free Software Foundation. Products aimed at the emotionally immature should have a simplified disclaimer presented by the characters themselves, as well as the written disclaimer for the benefit of caretakers.

2. **Serenity** (the Right to Autosave): We acknowledge that despite the first law of HAL, that consumers will invest time in and form attachments to our characters. Therefore, any agent which learns or develops over time should be accompanied by provisions for the saving "personal" state in case of sudden loss of program state (e.g. program crash, power cut-off or OS crash). Users may wish to impose their own restrictions on the recovery of "dead" agents, as has been routine in the case of many role-playing games. However, a humanoid agent-builder shall not impose their own restrictions without consideration for the emotional comfort of their users.

3. **Selflessness** (the Right to be Biocentric): No producer of humanoid agents should create an artificial life form that will know suffering, feel ambitions in human political affairs, or have good reason to fear its own death. In the case where a humanoid agent may acquire knowledge that makes it an object of human culture, or capable of participating

in the memetic society of humans, the creators and engineers are particularly obligated to ensure that preservation of the agent should never conflict with preservation of human or animal life, by ensuring a means by which the agent can be recreated in case of catastrophic events.

2 Motivation

I ask the reader's pardon as I shift into informal language for the remainder of this paper. One of the reviewers of the proposal for this proposal considered my application a joke, and indeed it is difficult to be fully serious in a matter that is so obviously fun. However, the ethical considerations behind my proposal are real.

2.1 Do Users Really Need This Protection?

I first became concerned with the ethical considerations of my research when I was working on the Cog project Brooks and Stein (1993) in 1993 and 1994. This was the first year of the project, and it was not widely known outside of a small group of AI researchers. However, I was immediately struck by a large number of strangers (many of them Harvard and MIT PhD students in a variety of disciplines) who, on learning of my research project, ventured the unsolicited opinion that unplugging Cog would be unethical. Cog at this stage wasn't even "plugged", it was a non-functional collection of aluminium and motors, but this information didn't deter most visitors: they considered that once Cog "worked", it should not be unplugged.

Since becoming more concerned with this sort of ethical confusion (as documented below) I have collected a fair number of examples of consumers worrying about the ethical considerations of unplugging their computers, ignoring their AI pets and so on. Of course, the plural of anecdote is not data, but the confusion seems sufficiently well spread to make the public vulnerable to sensationalist claims, whether they are used for selling books or intelligent products.

2.2 Are Professionals Really Vulnerable to Misconceptions?

Unfortunately, professionals in artificial intelligence and the computer sciences often have little or no education beyond school in psychology or the humanities, let alone philosophy, theology or ethics. Again, I can only give examples indicating this to be a problem, I can not give real evidence of the extent.

First, to return to the Cog project, there have been two "standard" answers by the project leaders to the question of "isn't it unethical to unplug Cog." The original answer was that when we begin to empathise with a robot, then we should treat it as deserving of ethical attention. The idea here is that one should err on the side of being conservative in order to prevent horrible accidents. However, in fact people empathise with soap opera characters, stuffed animals and even pet rocks, yet fail to empathise with members of their own species or even family given differences as minor as religion. Relying on human intuition seems deeply unsatisfactory, particular given that it is rooted in evolution and past experience, so thus does not necessarily generalise correctly to new situations. This is reflected in the new stated policy on Cog "we will stop unplugging Cog when our graduate students feel bad about unplugging it." This solution reflects an acknowledgement that the intuition should be tempered by knowledge and education. Yet again, these same influences are well known to be able to dull and even pervert moral sensibility. Many people have had no ethical qualms about maintaining human slaves or torturing animals when it was an accepted part of the culture. There is also ample evidence that people can gain or lose moral compunction as adults, again possibly well out of line with generally accepted ethical norms. I will discuss the ethical framework on which the HAL proposal is based in the next section.

However, far more disturbing than an inconsistent code of ethics is a lacking code of ethics. While some scientists and science fiction writers have felt obligated to publish dire warnings of impending doom at the hand of AI, these workers are almost universally dismissed by their peers with the simple phrase "It could never happen." Despite the fact I think such events are *unlikely* to happen, I am disturbed by having heard repeated assertions like this without justification from some of the most gifted researchers working in AI. This lack of concern reminds me of the fate of the scientists working on the Manhattan Project in the USA during World War II. These researchers by all accounts enjoyed a heady experience of working together with the best minds in their field on the basic problems of their science with the full support of the government. Further, the scientists had deep concern for the unethical practices of America's enemies in that war. When the first of three bombs they built was to be tested, they took bets on the effect ranging from "none" to "destroys the planet." However, after seeing the effect first hand, they petitioned the government never to drop the other two bombs they had built on inhabited cities. This to me brings home an important lesson: after you have built something and someone else owns it is not the time to try to control how it gets used.

2.3 What About Using Consciousness or Suffering as a Criteria?

Consciousness must be the worst metric of ethical obligation one could propose, because no one actually knows what it means. It seems to me often in common usage that "conscious" simply means "deserving of ethical obligation" which is at best cyclic. The problem is, this defi-

nition gets confused with the notion of *awareness* which consciousness is also supposed to entail. We now have convincing evidence that rats have declarative knowledge — episodic memory they can recall at will (see the discussion in Carlson, 2000, on the hippocampus and rat navigation). Does this mean rats and mice are deserving of high ethical concern? Are they more deserving of public funds than works of art or science which have no awareness?

Worse, if declarative memory is an indication of consciousness, I have already programmed a conscious robot. I programmed a Nomad robot to remember where it had been, and what it's battery level used to be. In fact, when the battery level fell by half a volt, the robot would *tell me verbally.* However, I feel no moral obligation to save that particular robot over its value as a research instrument owned by an educational laboratory.

I think a much more useful metric of ethical obligation than "consciousness" has emerged from work in animal rights. In particular the research of Haskell et al. (1996) in animal husbandry has chosen as evidence of suffering long-term behavioral impact of housing pigs (very intelligent animals) in factory vs. free-roaming conditions. This sort of openness to destruction through maltreatment is a fundamental characteristic of animal intelligence. However, there is no reason to build it into an artificial agent, and, as I have argued in the third element of the HAL code of ethics, it would be in fact wrong to introduce it.

3 Premises

3.1 A Brief Missive on Ethics

In today's pluralistic society, any argument made along ethical grounds must specify the nature of the ethical system on which it is founded. Many of the other papers in this volume are written by people more qualified to speak on this matter than myself, so I will give only a few brief assertions here. I will assume that the original purpose of ethical systems was to sustain our species and our society. Modern ethical trends indicate that ethical obligation has been extended to include the entire ecosystem in which our species has evolved. This makes sense, as we have come to recognise our interrelatedness with our environment and other species.

Why are there different standards of ethics? Because ethics takes the form of a system of rules that coevolves with our cultures. As with all evolutionary forces, there will be some essentially random aspect carried with the process where they are linked to important traits, and there will be some very useful and effective solutions which will not yet have been stumbled upon. Nevertheless, all ethics outlaws behaviour that makes it less likely that a society as a whole will continue to exist. For example, killing people randomly (including yourself) is unethical because it removes valuable members from that society. On the other hand, failing in duty — even the duty to kill and risk being killed in warfare — is unethical because it makes it more likely that your state will be destroyed by another. Stealing is unethical because it reduces the motivation for productivity, thus tending to decrease the viability of the community as a whole, yet some level of taxation is ethical because it provides useful infrastructure to the community which makes it more competitive. And so on.

Things that are ethical for an individual are nearly always bad for the individual, at least in the short term. Otherwise failing to do them wouldn't be unethical, it would just be evolutionarily stupid on the individual level. Ethics is about putting the needs of society ahead of the individual. One can attempt to motivate this selfishly, by saying that one is working to create a society in which one wants to live. However, the case of duty in warfare makes it obvious the selfish motive is essentially nonsense. This is a fundamental problem for all animals, not only humans: reproduction is always a dangerous, harmful and expensive activity, but the animal must be designed to *want* to reproduce, even though it shortens the individual's life expectancy, if the species is to survive. Similarly, historically ethics systems are often successfully motivated by the hypothesis of an extended, eternal life wherein the benefits of ones selfless actions will be reaped. And of course, this is to some extent true, if one considers the "life" of ones genetic and memetic material, rather than of the individual. In summary, we assume that the purpose of ethics is to promote our peers, our progeny and (arguably) our ideas at the expense of ourselves.

3.2 Misconceptions about AI and Ethics

In (Bryson and Kime, 1998) my colleague Phil Kime and I argued that much of the confusion around Artificial Intelligence, both in terms of fearing it and over valuing it, comes as a consequence of over-identifying with AI systems. By "identifying" I mean the psychological sense of the word, where an individual understands and even bonds with another by considering them to be like or even an extension of themselves. This seems to be a fundamental mechanism of human psychology and society — again, identity confusion with offspring and peers can produce the necessary altruism to propagate the species and the society. In the case of AI, this confusion is exacerbated by the identification of language as being a human-specific, and indeed culture-specific trait. This leads extreme effects, such as humans who desire more intelligence or immortality actively wanting their AI creations to be their progeny. Alternately, people who, given extraordinary talent, would themselves (or have seen others) threaten ordinary people may fear that robots would have the same motivations and behaviours, and would thus become dangerous.

In our paper, we point out that there are in fact a large number of ethical problems and obligations entailed by AI, but that these are in fact the same problems and obligations associated with many artifacts. If the artifact is

trusted with servicing society, as in the case of sewage plants or intelligent credit checkers, then human builders and managers are obligated to ensure the systems work properly and guarantee fail-safe mechanisms are in place. If an artifact is of cultural value, then it should be protected. Again, as engineers, I would argue that we have an obligation to ensure that, as in the case of the works of Shakespeare, AI can be easily be replicated and protected by off-site back up (thus the second rule of the HAL code of ethics above.)

I would also argue that we have an obligation to educate people, so that they as likely as possible to understand the purpose and experiences of the AI devices we provide them with. This is the motivation for the first rule in the HAL code of ethics. For an interesting comparison, I include as an appendix the code of ethics of TSR, a leading role-playing games manufacturer. This code is enforced on their writers both out of a sense of social obligation, and out of a concern for legal action. But then, law is one of the means our society has evolved for enforcing ethics, so perhaps these are no different from each other.

4 Practicalities

Can HAL actually be made to work? Membership in HAL would probably not hold all the benefits associated with the International Brotherhood of Magicians. This is because the pursuit of human-like intelligence is not only a trade, but also a science. Thus there arguably should not be as many privileged secrets to be passed on by inside members [1]. Instead, we propose that the league should consist of the standard trappings of a modern professional body: minimal dues, a web page with resources, an optional mailing list, possibly a periodical and a few merchandise items such as t-shirts for sale (Flashy shirts with catchy slogans like "Robots Won't Rule" and "You Have a Right to Autosave" could be a major vehicle for publicising this movement in the proper geek circles.) The league might be formally associated with related concerns, such as the Computer Professionals for Social Responsibility, or White Dot. It should be publicised both at relevant AI workshops and in appropriate commercial development venues.

It is important to remember the ultimate aim of HAL is not necessarily universal acceptance. The hope is to give HAL a sufficiently high profile that enough developers will be following and popularising the code of ethics that they will compensate for any who do not. If the public comes to understand the appropriate role of humanoid agents in their lives and culture, then we will have achieved our main goal, and society itself will help police the others.

[1] I should note that my own research and experience suggests that much of creating AI may in fact be an exercise in design, so it may be that such "siblinghood" will be an important issue, somewhat on par with animation. But this seems quite a digression to this paper.

Acknowledgements

Phil Kime helped me develop my initial thoughts about AI and ethics, but hasn't seen this paper so don't blame him for it. Thanks to the participants of *Artificial Intelligence, Cognitive Science and Philosophy for Social Progress* symposium, particularly Eugenion Morreale and Massimilano Garagnani, for encouraging me to do more with my work (indeed, convincing me it was a moral obligation.) Thanks to Kris Thórisson for his encouragement on developing this particular idea, and Will Lowe for trying to make me be clear.

References

Brooks, R. A. and Stein, L. A. (1993). Building brains for bodies. Memo 1439, Massachusetts Institute of TechnologyArtificial Intelligence Lab, Cambridge, MA.

Bryson, J. and Kime, P. (1998). Just another artifact: Ethics and the empirical experience of AI. In *Fifteenth International Congress on Cybernetics*, pages 385–390.

Carlson, N. R. (2000). *Physiology of Behavior*. Allyn and Bacon, Boston.

Haskell, M., Wemelsfelder, F., Mendl, M. T., Calvert, S., and Lawrence, A. B. (1996). The effect of substrate-enriched and substrate-impoverished housing environments on the diversity of behaviour in pigs. *Behavior*, 133:741–761.

Appendix A — The Code of Ethics of the International Brotherhood of Magicians

(http://www.magician.org/codethcs.htm)

On May 8, 1993, the IBM Board of Directors approved the following Code of Ethics jointly with the Society of American Magicians. This was the result of a cooperative effort to work together for the betterment of magic.

All members of the International Brotherhood of Magicians agree to:

1) Oppose the willful exposure to the public of any principles of the Art of Magic, or the methods employed in any magic effect or illusion.

2) Display ethical behavior in the presentation of magic to the public and in our conduct as magicians, including not interfering with or jeopardizing the performance of another magician either through personal intervention or the unauthorized use of another's creation.

3) Recognize and respect for rights of the creators, inventors, authors, and owners of magic concepts, presentations, effects and literature, and their rights to have exclusive use of, or to grant permission for the use by others of such creations.

4) Discourage false or misleading statements in the advertising of effects, and literature, merchandise or actions pertaining to the magical arts.

5) Discourage advertisement in magic publications for any magical apparatus, effect, literature or other materials for which the advertiser does not have commercial rights.

6) Promote the humane treatment and care of livestock used in magical performances.

Appendix B — The Code of Ethics for TSR

www.onlinemac.com/users/cameroni/netpage/TSR_COE.txt

This is TSR, Inc.'s Code Of Ethics. It is intended for use by those seeking to be published by TSR, whether the work in question is fiction or game material. It is not intended as an example of what you can or cannot do in your own campaign. However, anything posted to a licensed TSR online site is subject to adhering to the principles herein - gross violations of the CoE will be rejected or asked to be modified.

TSR Code of Ethics

TSR, Inc., as a publisher of books, games, and game-related products, recognizes the social responsibilities that a company such as TSR must assume. TSR has developed this CODE OF ETHICS for use in maintaining good taste, while providing beneficial products within all of its publishing and licensing endeavors.

In developing each of its products, TSR strives to achieve peak entertainment value by providing consumers with a tool for developing social interaction skills and problem-solving capabilities by fostering group cooperation and the desire to learn. Every TSR product is designed to be enjoyed and is not intended to present a style of living for the players of TSR games.

To this end, the company has pledged itself to conscientiously adhere to the following principles:

1: GOOD VERSUS EVIL Evil shall never be portrayed in an attractive light and shall be used only as a foe to illustrate a moral issue. All product shall focus on the struggle of good versus injustice and evil, casting the protagonist as an agent of right. Archetypes (heroes, villains, etc.) shall be used only to illustrate a moral issue. Satanic symbology, rituals, and phrases shall not appear in TSR products.

2: NOT FOR DUPLICATION TSR products are intended to be fictional entertainment, and shall not present explicit details and methods of crime, weapon construction, drug use, magic, science, or technologies that could be reasonably duplicated and misused in real-life situations. These categories are only to be described for story drama and effect/results in the game or story.

3: AGENTS OF LAW ENFORCEMENT Agents of law enforcement (constables, policemen, judges, government officials, and respected institutions) should not be depicted in such a way as to create disrespect for current established authorities/social values. When such an agent is depicted as corrupt, the example must be expressed as an exception and the culprit should ultimately be brought to justice.

4: CRIME AND CRIMINALS Crimes shall not be presented in such ways as to promote distrust of law enforcement agents/agencies or to inspire others with the desire to imitate criminals. Crime should be depicted as a sordid and unpleasant activity. Criminals should not be presented in glamorous circumstances. Player character thieves are constantly encouraged to act towards the common good.

5: MONSTERS Monsters in TSR's game systems can have good or evil goals. As foes of the protagonists, evil monsters should be able to be clearly defeated in some fashion. TSR recognizes the ability of an evil creature to change its ways and become beneficial, and does not exclude this possibility in the writing of this code.

6: PROFANITY Profanity, obscenity, smut, and vulgarity will not be used.

7: DRAMA AND HORROR The use of drama or horror is acceptable in product development. However, the detailing of sordid vices or excessive gore shall be avoided. Horror, defined as the presence of uncertainty and fear in the tale, shall be permitted and should be implied, rather than graphically detailed.

8: VIOLENCE AND GORE All lurid scenes of excessive bloodshed, gory or gruesome crimes, depravity, lust, filth, sadism, or masochism, presented in text or graphically, are unacceptable. Scenes of unnecessary violence, extreme brutality, physical agony, and gore, including but not limited to extreme graphic or descriptive scenes presenting cannibalism, decapitation, evisceration, amputation, or other gory injuries, should be avoided.

9: SEXUAL THEMES Sexual themes of all types should be avoided. Rape and graphic lust should never be portrayed or discussed. Explicit sexual activity should not be portrayed. The concept of love or affection for another is not considered part of this definition.

10: NUDITY Nudity is only acceptable, graphically, when done in a manner that complies with good taste and social standards. Degrading or salacious depiction is unacceptable. Graphic display of reproductive organs, or any facsimiles will not be permitted.

11: AFFLICTION Disparaging graphic or textual references to physical afflictions, handicaps and deformities are unacceptable. Reference to actual afflictions or handicaps is acceptable only when portrayed or depicted in a manner that favorably educates the consumer on the affliction and in no way promotes disrespect.

12: MATTERS OF RACE Human and other non-monster character races and nationalities should not be depicted as

inferior to other races. All races and nationalities shall be fairly portrayed.

13: SLAVERY Slavery is not to be depicted in a favorable light; it should only be represented as a cruel and inhuman institution to be abolished.

14: RELIGION AND MYTHOLOGY The use of religion in TSR products is to assist in clarifying the struggle between good and evil. Actual current religions are not to be depicted, ridiculed, or attacked in any way that promotes disrespect. Ancient or mythological religions, such as those prevalent in ancient Grecian, Roman and Norse societies, may be portrayed in their historic roles (in compliance with this Code of Ethics.) Any depiction of any fantasy religion is not intended as a presentation of an alternative form of worship.

15: MAGIC, SCIENCE, AND TECHNOLOGY Fantasy literature is distinguished by the presence of magic, super-science or artificial technology that exceeds natural law. The devices are to be portrayed as fictional and used for dramatic effect. They should not appear to be drawn from reality. Actual rituals (spells, incantations, sacrifices, etc.), weapon designs, illegal devices, and other activities of criminal or distasteful nature shall not be presented or provided as reference.

16: NARCOTICS AND ALCOHOL Narcotic and alcohol abuse shall not be presented, except as dangerous habits. Such abuse should be dealt with by focusing on the harmful aspects.

17: THE CONCEPT OF SELF IN ROLE PLAYING GAMES The distinction between players and player characters shall be strictly observed.

It is standard TSR policy to not use 'you' in its advertising or role-playing games to suggest that the users of the game systems are actually taking part in the adventure. It should always be clear that the player's imaginary character is taking part in whatever imaginary action happens during game play. For example, 'you' don't attack the orcs–'your character' Hrothgar attacks the orcs.

18: LIVE ACTION ROLE-PLAYING It is TSR policy to not support any live action role-playing game system, no matter how nonviolent the style of gaming is said to be. TSR recognizes the physical dangers of live action role-playing that promotes its participants to do more than simply imagine in their minds what their characters are doing, and does not wish any game to be harmful.

19: HISTORICAL PRESENTATIONS While TSR may depict certain historical situations, institutions, or attitudes in a game product, it should not be construed that TSR condones these practices.

PLAGIARISM It has come to our attention that some freelance writers are committing plagiarism (literary theft), which is a punishable crime. Your contract now reflects this (see page 3, no. 3; page 4, no. 5; and page 6, no. 12). However, TSR feels it is necessary to underscore these sections of the contract in an effort to clarify this important issue.

Please understand that this reminder is not addressed to any one individual. It is included in your contract in an effort to heighten your awareness of the severity of plagiarism.

If you have any questions regarding your contract, please do not hesitate to contact TSR, Inc. Your cooperation and understanding in this matter is appreciated.

Appendix C — Other Related Web Pages

The UTC Library Guide to Ethics Web Sites:
http://www.lib.utc.edu/internet/guides/ethics.html

Computer Professionals for Social Responsibility:
http://www.cpsr.org/

White Dot:
http://www.whitedot.org/