WHAT IS EMPATHETIC SUPERINTELLIGENCE?

Overview

Our current conception of intelligence as measured by IQ tests is “mind-blind”. IQ tests lack ecological validity because they ignore social cognition – the “mind-reading” prowess that enabled one species of social primate to become the most cognitively successful on the planet. In this talk, I shall examine how to correct the ethnocentric and anthropocentric biases of our perspective-taking abilities. What future technologies can enrich our capacity to understand other minds? I shall also discuss obstacles to building empathetic AGI (artificial general intelligence) - and why old-fashioned “autistic AI” may always be vulnerable to the cunning of “Machiavellian apes”. In an era of global catastrophic and existential risks, developing ways to enrich and “de-bias” mankind's capacity for empathetic cognition will be vital. This is because the greatest underlying risk to the well-being of life on Earth is the dominance behaviour of other male human primates.

INTRODUCTION

Q: What is our greatest underlying source of:

1. global catastrophic risk and existential risk?
2. violence, war, oppression of other ethnic groups and species?
3. low mood, social anxiety and clinical depression? (*cf.* Rank Theory)

INTRODUCTION

A: the competitive male dominance behaviour of Machiavellian human primates.

An evolutionary spiral of "mind-reading" prowess has helped one species of social primate become the most cognitively successful on the planet. But "Machiavellian intelligence" is biased, partial and selective. *Egocentric bias, ethnocentric bias and anthropocentric bias* were fitness-enhancing in the ancestral environment of adaptation. Hence the horrors of the slave trade, Auschwitz - and contemporary agribusiness. Humans must become both better "systematizers" and better "empathizers".

PROPOSAL:

Only biological remediation ("enhancement") can cure our profound genetic deficits of empathetic understanding: We need to "de-bias" social intelligence and achieve an impartial "God's-eye-view" in science and ethics alike.

Refs:

1. Premack, D. G. & Woodruff, G. (1978). "Does the chimpanzee have a theory of mind?" *Behavioral and Brain Sciences*, 1, 515-526.
2. Baron-Cohen S (2009). "Autism: the empathizing–systemizing (E-S) theory". *Ann N Y Acad Sci* 1156: 68–80
3. "The relationship between empathy and Machiavellianism: An alternative to empathizing–systemizing theory" J. Andrewa, M. Cookea and S.J. Muncer. *Personality and Individual Differences* Volume 44, Issue 5, April 2008, Pages 1203-1211
4. http://en.wikipedia.org/wiki/Rank_theory_of_depression
5. Wrangham, R. and Peterson, D. *Demonic Males* (1996).
6. *Chimpanzee Politics: Power and Sex among Apes* (2000, 2007) by Frans de Waal
7. *Global Catastrophic Risks* (2008)
   Nick Bostrom (Ed), Milan M. Cirkovic (Ed)

SUPER-JAINISM?

Jain monks and nuns:

- believe in the sanctity of life
- practise *ahimsa*, "harmlessness"
- are vegans
- walk barefoot
- sweep the ground in front of them with a whisk broom to avoid killing insects, worms and other small creatures.

PARALLEL?

- *Homo sapiens* and small insects
- Posthumans and humans

Should we aspire to become Super-Jains and brush - literally and figuratively - the path in front of us?

SUPER-JAINISM?

ULTIMATELY, **YES**.

BUT….To be effective, benevolence must be intelligent, rational and systematic. A God-like superintelligence would impartially weigh all possible first-person perspectives - from simple desires to higher-order intentionality, the world-simulations of humble bumblebees and massive Jupiter brains.

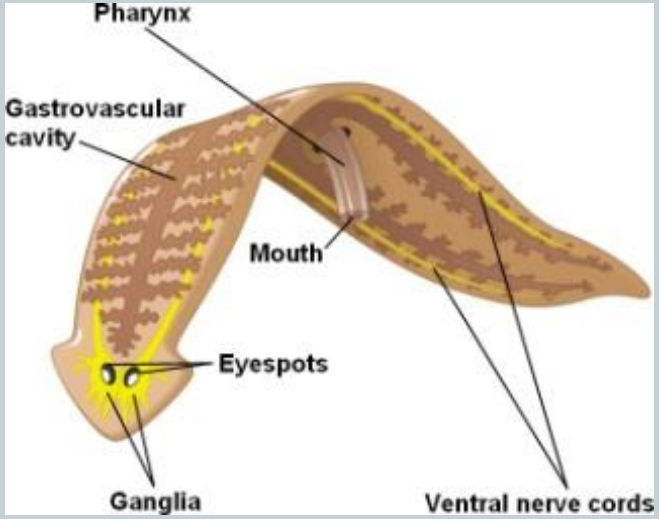
Refs:
1. http://en.wikipedia.org/wiki/Jainism
2. http://en.wikipedia.org/wiki/Theory_of_mind

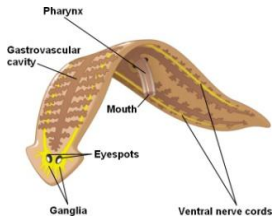
THE FUTURE OF SENTIENCE

"It seems plausible that with technology we can, in the fairly near future create (or become) creatures who surpass humans in every intellectual and creative dimension. Events beyond such an event — such a singularity — are as unimaginable to us as opera is to a flatworm."
Vernor Vinge

What do opera-loving humans have in common with flatworms - and not with classical symbolic AI systems and paperclips?

- Sentience
- An opioid-dopamine system / a pleasure-pain axis
- Without the phenomenology of the pleasure-pain axis, nothing *inherently* matters





THE FUTURE OF SENTIENCE

QUESTIONS:

1) Why aren't organic robots mere zombies? (the Hard Problem of Consciousness, the Explanatory Gap)

2) Does the future of life lie in super-AGIs governed by formal utility functions - or in reinforcement learning, gradients of intelligent bliss and mind-reading organic robots?

   Or hybrids of both?

   Or neither?

Refs:
1. http://plato.stanford.edu/entries/zombies/
2. http://en.wikipedia.org/wiki/Hard_problem_of_consciousness
3. http://en.wikipedia.org/wiki/Explanatory_gap
4. Galen Strawson (2006). Realistic Monism - Why Physicalism Entails Panpsychism. *Journal of Consciousness Studies* 13 (10-11):3-31
5. *The Singularity Is Near: When Humans Transcend Biology* (2005) by Ray Kurzweil

WHAT IS FRIENDLINESS?

"Until he extends the circle of his compassion to all living things, man will not himself find peace."
Albert Schweitzer
(The Philosophy of Civilisation, Tr. by C.T. Campion, New York, Macmillan Co., 1949.)

"Our task must be to free ourselves... by widening our circle of compassion to embrace all living creatures and the whole of nature and its beauty."
Albert Einstein

ASSUMPTION: Minimum precondition of any advanced civilisation:

"We advocate the well-being of all sentience, including humans, non-human animals, and any future artificial intellects, modified life forms, or other intelligences to which technological and scientific advance may give rise"
(Transhumanist Declaration 1998, 2009)

but...

WHAT IS FRIENDLINESS?

Well-being: Questions

- What kind of well-being?
- How far above “hedonic zero”?
- How extensive? [world-wide, pan-galactic, cosmic?]
- Traditional meat world or Virtual Reality?
- Egoistic or Empathetic?
- Solitary or Social?
- Intelligent or Orgasmic?
- Plural or singleton?
- Jupiter brains and/or a utilitronium shockwave in the maximum density of Everett branches?

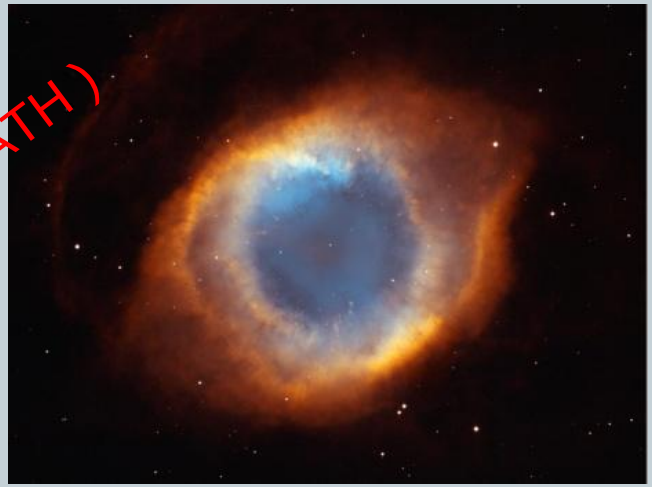
Further Questions:

- What should we do about predators, sociopaths, malevolent agents?
- Is friendliness computable?

Refs:
1. http://humanityplus.org/learn/transhumanist-declaration/
2. http://www.abolitionist.com/reprogramming/
3. http://www.abolitionist.com/multiverse.html

A GOD'S-EYE-VIEW

Should we aim for *human*-friendly superintelligence i.e. to "lock in" anthropocentric bias?
Or should we aspire to *sentience*-friendly superintelligence?



An impartial "view from nowhere" embraces the state-space of all possible minds (*cf.* the timeless Wheeler-DeWitt equation in quantum cosmology)

QUESTIONS:
Could a benevolent SuperIntelligence exhibit status quo bias?
Will posthumans share our naive conceptions of personal identity?
(*cf.* Buddhist, ultra-Parfitian[1] conceptions of personal (non-)identity versus an enduring metaphysical ego)

Why Does Our Theory of Personal Identity Matter?
Retarded Maturation versus Accelerated Maturation
ANALOGY: Are adult humans "baby-friendly"? We aim to replace babies with adults via education. But we aren't "baby-killers." Adult caregivers help babies "grow up".
Likewise, how fast - and how far - might posthumans educate or "uplift" simple-minded human primates into mature posthumans?
Ultimately, how much of Darwinian life is worth perpetuating or (re)creating?

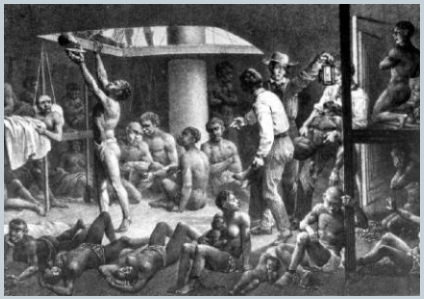
refs:
1. *Reasons and Persons* (1984) by Derek Parfit
2. *The View from Nowhere* (1986) by Thomas Nagel
3. *The End of Time* (1999) by Julian Barbour

THE EVOLUTIONARY ORIGINS OF UNFRIENDLINESS IN ORGANIC ROBOTS

Ultimate and Proximate Causes

"The history of the world my sweet, is who gets eaten and who gets to eat..."
(*Sweeney Todd: The Demon Barber of Fleet Street* (1979))

- "selfish DNA" - the evolution of information-bearing self-replicators
- Natural selection. Sexual reproduction. The Cambrian explosion. The "trophic levels" of the food chain.
- Darwin. Mendel. Modern Evolutionary Synthesis.
- Ultra-Darwinism, The Extended Phenotype (Dawkins)
- the "encephalisation of emotion"
- Rise of *Homo sapiens*



"History is indeed little more than the register of the crimes, follies, and misfortunes of mankind."
Edward Gibbon
(*The History Of The Decline And Fall Of The Roman Empire* (1776))

QUESTION: Would true friendliness necessitate 100% genetic identity?

refs:
1. *Adaptation and Natural Selection* (1966) by George C. Williams
2. *The Selfish Gene* (1976) by Richard Dawkins
3. *Genes in Conflict: The Biology of Selfish Genetic Elements* (2006) by Austin Burt
4. http://en.wikipedia.org/wiki/Eusocial

WHY ARE HUMANS THE MOST COGNITIVELY SUCCESSFUL SPECIES ON EARTH? WHAT IS INTELLIGENCE?

Do existing IQ tests have ecological validity?
Current IQ tests are “mind-blind”. They measure only autistic intelligence.

OBJECTION
Isn't “empathetic intelligence” really just the personality variable of agreeableness? Simon Baron-Cohen contrasts the “autistic” cognitive style with the “empathetic” cognitive style. The autistic cognitive style searches for truth - defined as precise, reliable, consistent, or lawful patterns or structure in data. By contrast, the empathetic cognitive style - focused on caring, nurturing - was evolutionarily advantageous for female caregivers.

WHY ARE HUMANS THE MOST COGNITIVELY SUCCESSFUL SPECIES ON EARTH? WHAT IS INTELLIGENCE?

POSSIBLE ANSWER

No. Our mind-reading capacity is a cognitively demanding adaptation: mind-reading sometimes demands fifth-order or even sixth-order intentionality. ["I think that you hope that she believes that I want you to think he desires to...."] But our perspective-taking skills have been systematically warped by evolution to maximise the inclusive fitness of our genes: "Machiavellian intelligence" facilitates tribal hunting, genocide, "big science", factory-farming, the Manhattan Project, Microsoft and the infrastructure of modern society. A *partial* and *selective* empathetic understanding of the social world e.g. division into "allies" and "rivals", "us" and "them", also makes for superior interrogators, torturers, military intelligence officers ("etc.")

"Imperfect" empathetic intelligence is dangerous.

Refs:
1. Frans de Waal's *Chimpanzee Politics* (1982)
2. Testosterone decreases trust in socially naive humans. Bos PA, Terburg D, van Honk J.
Proc Natl Acad Sci U S A. 2010 Jun 1;107(22):9991-5.
3. Flynn, James R. *What Is Intelligence: Beyond the Flynn Effect* (2009)
4. Gardner, Howard (1983; 1993) *Frames of Mind: The Theory of Multiple Intelligences*
5. Evolution of spite through indirect reciprocity. Rufus A Johnstone and Redouan Bshary
Proc Biol Sci. 2004 September 22; 271(1551): 1917–1922.
6. Dunbar, R.I.M. (1993), Coevolution of neocortical size, group size and language in humans,
Behavioral and Brain Sciences 16 (4): 681-735.
7. The relationship between empathy and Machiavellianism: An alternative to empathizing–systemizing theory
J. Andrewa, M. Cookea and S.J. Muncer
Personality and Individual Differences Volume 44, Issue 5, April 2008, Pages 1203-1211

ROUTES TO BIO-HAPPINESS AND BIO-FRIENDLINESS

Bio-happiness and bio-friendliness compared:

1) Wireheading / neurostimulation
(*cf*. José Delgado's experiments on raging bulls)
2) Euphoriant drugs / empathetic drugs
(can they recalibrate the hedonic treadmill?
upregulate the oxytocinergic system in the CNS?)
3) Genetic recalibration of our "hedonic set point" / genetically enhanced predisposition to empathetic understanding

Are genetically preprogrammed gradients of superhappiness easier to hardwire than genetically preprogrammed superfriendliness?

Human nature means social reform, on its own, can't abolish suffering, let alone make us superhappy (*cf*. the hedonic treadmill)
Can social engineering ever make competitive male human primates safe and friendly, let alone superfriendly?
Will posthumans regard their ancestors as quasi-sociopaths?

Refs:
1. *Physical Control of the Mind: Toward a Psychocivilized Society* (1969) by José M. R. Delgado, M.D.
http://www.wireheading.com/delgado
2. Investigating the genetic basis of altruism: the role of the COMT Val158Met polymorphism Martin Reuter, Clemens Frenzel, Nora T. Walter, Sebastian Markett, and Christian Montag. *Soc Cogn Affect Neurosci* (2010) doi: 10.1093/scan/nsq083
3. The Catechol-O-Methyl Transferase Val158Met Polymorphism and Experience of Reward in the Flow of Daily Life
Marieke Wichers, Mari Aguilera, Gunter Kenis, Lydia Krabbendam, Inez Myin-Germeys, Nele Jacobs, Frenk Peeters, Catherine Derom, Robert Vlietinck, Ron Mengelers, Philippe Delespaul and Jim van Os. *Neuropsychopharmacology* (2008) 33, 3030–3036
4. Oxytocin promotes human ethnocentrism
Carsten K. W. De Dreu1, Lindred L. Greer, Gerben A. Van Kleef, Shaul Shalvi, and Michel J. J. Handgraaf *PNAS* January 25, 2011 vol. 108 no. 4 1262-1266

WHAT'S THE EASIEST WAY TO DISARM A POTENTIALLY HOSTILE ALPHA MALE ORGANIC ROBOT?

1. Give him a good education, civics classes, a course of utilitarian ethics / virtue theory / deontological ethics, or whatever?

WHAT'S THE EASIEST WAY TO DISARM A POTENTIALLY HOSTILE ALPHA MALE ORGANIC ROBOT?

2. Give him a generous dose of an empathogen and get smothered with hugs…

CASE STUDY:
MDMA ("Ecstasy") – "the penicillin of the soul"?

Dopamine release - pleasure
Serotonin release - heightened emotion
Oxytocin release - trust, bonding, honesty, self-love

"I love the world and the world loves me"

Until banned, the drug was used by US psychotherapists as a therapeutic tool: MDMA induces unparalleled emotional honesty, trust and self-disclosure.

Refs:
1. *Neuroscience*. 2007 May 11;146(2):509-14. Epub 2007 Mar 23.
A role for oxytocin and 5-HT(1A) receptors in the prosocial effects of 3,4 methylenedioxymethamphetamine ("ecstasy").
Thompson MR, Callaghan PD, Hunt GE, Cornish JL, McGregor IS.
2. *Eur J Neurosci*. 2004 Aug;20(3):853-8.
The rewarding properties of MDMA are preserved in mice lacking mu-opioid receptors.
Robledo P, Mendizabal V, Ortuño J, de la Torre R, Kieffer BL, Maldonado R.
3. Dennett, D. "Three kinds of intentional psychology" (IP) in Heil, J. - Philosophy of Mind: A guide and anthology, (2004)
4. *Ecstasy: The Complete Guide* (2001). by Julie Holland, M.D
5. http://www.mdma.net

THE RAVERS' MOTTO : P.L.U.R.

THE RAVERS' MOTTO : P.L.U.R.
Peace, Love, Unity and Respect

Peace - Letting go of fear and living at peace with oneself, one another, and the planet for a greater good.

Love - As one learns to love oneself, one is able to love everyone else unconditionally.

Unity - A mutual, corporate bond is formed, resulting from the love and peace experienced with one another.

Respect - Because of peace, love and unity, one can accept others regardless of their beliefs or background

Ref:
(from Hyperreal, Netraver, etc)

The Problem of (Un)Friendliness
Why Empathogenic Drugs Are (probably) Not The Answer

PITFALLS

- Short-acting
- Activates multiple negative feedback mechanisms
- Heavy MDMA users may be less empathetic, less friendly, more irritable "on the rebound"
- "Abuse potential"
- Not conducive to economic growth

Et cetera

BUT:

Could safe, sustainable **oxytocin** analogues promote friendliness with minimal abuse potential?

(*cf.* the positive correlation between level of trust in different societies and economic growth)

Do we want to design safe and sustainable euphoriants and empathogens for fine-grained, reversible modulation of mood and sociability?

Refs:
1. The oxytocin receptor (OXTR) contributes to prosocial fund allocations in the dictator game and the social value orientations task. *PLoS One*. 2009 May 20;4(5): e5535.
2. Oxytocin: Crossing the Bridge between Basic Science and Pharmacotherapy
CNS Neurosci Ther. 2010 October; 16(5): e138–e156.
doi: 10.1111/j.1755-5949.2010.00185.x.
3. Receptor and behavioral pharmacology of WAY-267464, a non-peptide oxytocin receptor agonist.
Neuropharmacology. 2010 Jan;58(1):69-77.

QUESTION:
Could The Functional Equivalent Of An Empathogenic "Hug Drug" Tame A Potentially Non-Friendly Super-AGI?

TENTATIVE ANSWER

Maybe. If our root-metaphor of intelligent mind derives from digital computers and symbolic AI, then probably not. Friendliness will be the outcome of inspired programming by the creators of the primordial seed AI. But an autistic AGI incapable of taking the intentional stance would be crippled intelligence. If, on the other hand, our root-metaphor of intelligent mind comes from connectionist neural networks, dynamical systems theory, or quantum mind (etc), then the functional equivalent of perpetual oxytocin flooding is more plausible.

BUT

Isn't a predisposition to super-friendless best genetically hardwired into both organic robots and AGI?

[implementation details omitted; margin too small]

AND

Might a "hug" from a Benevolent Super-AGI with a classical utilitarian ethic convert us into utilitronium?

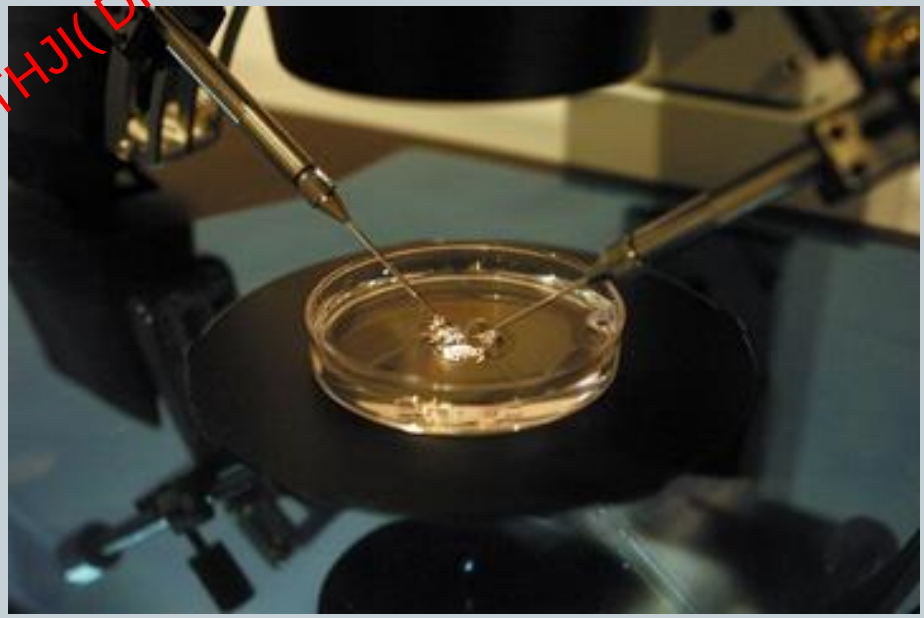
What happens if empathetic Superintelligence goes FOOM?

Ref:
1. http://singinst.org/
2. *The Intentional Stance* (1987, 1996) By Daniel C. Dennett

LONG-TERM SOLUTIONS: THE REPRODUCTIVE REVOLUTION

"Homo sapiens, the first truly free species, is about to decommission natural selection, the force that made us.... Soon we must look deep within ourselves and decide what we wish to become."
Edward O. Wilson (Consilience, The Unity of Knowledge 1998)

"The arrival of safe, reliable germline technology will [...] transform the evolutionary process by drawing reproduction into a highly selective social process that is far more rapid and effective at spreading successful genes than traditional sexual competition and mate selection."
Gregory Stock (Redesigning Humans (2002))



LONG-TERM SOLUTIONS: THE REPRODUCTIVE REVOLUTION

Proposal: future humans should innately happy and friendly, rather than discontented and distrustful.

DESIGNER BABIES

Prospective parents will shortly be able to pre-select the approximate hedonic set-point and empathetic intelligence of their future children via pre-implantation genetic diagnosis

Prospective parents will have the opportunity to:

1) boost lifelong oxytocin function (e.g. choose the "high empathy" G/G allele of the OXTR gene)

2) amplify mirror neurons ("the neurons that shaped civilisation")

3) genetically hardwire mirror-touch synaesthesia in their future children

Mirror-touch synaesthetes are hyper-empathetic compared to people without synaesthesia.

"I have never been able to understand how people can enjoy looking at bloodthirsty films, or laugh at the painful misfortunes of others when I can not only not look but also feel it."

(Alice, a mirror-touch synaesthete)

Refs:

1. Bakermans-Kranenburg MJ, van Ijzendoorn MH. Oxytocin receptor (OXTR) and serotonin transporter (5-HTT) genes associated with observed parenting. *Soc Cogn Affect Neurosci.* 2008 Jun;3(2):128-34.
2. *Annu. Rev. Neurosci.* 2004. 27:169–924 The Mirror-Neuron System Giacomo Rizzolatti and Laila Craighero
3. Banissy, M. J. & Ward, J. (2007). Mirror-touch synesthesia is linked with empathy. *Nature Neurosci.* doi: 10.1038/nn1926.
4. http://www.reproductive-revolution.com

Conclusion

Is life destined for:
An Empathy Explosion?
An autistic intelligence explosion?
An empathetic intelligence explosion?
All of the above?

TODAY, EMPATHETIC INTELLIGENCE ENTAILS SHARING THE SORROWS OF OTHER SENTIENT BEINGS. IN OUR POSTHUMAN FUTURE, WILL EMPATHY CONSIST ENTIRELY IN SHARING EACH OTHER'S JOYS?

THE END