Meaning, Leverage, and the Future

A Movement in Three Parts

Paris

6 may 2014

Richard Mallah

rim29@columbia.edu

Who I Am

- Lead AI R&D at semantic middleware company Cambridge Semantics
- Background in machine learning, computational linguistics, and knowledge representation
- Experience leading tech startups using narrow AI in various industries
- Led enterprise risk management systems at largest asset manager by AUM
- Personal interest and research in knowledge representation for AGI, maximizing benefits while minimizing risks
- Representative of the Future of Life Institute

The Future (of Life)
Existing organizations

- Centre for the Study of Existential Risk (CSER)
- Future of Humanity Institute (FHI)
- Machine Intelligence Research Institute (MIRI)
- Global Catastrophic Risk Institute (GCRI)
- **Less focused:** Union of Concerned Scientists, Federation of American Scientists, Foresight Institute, JASON defense advisory group
- **Less active:** Center for Responsible Nanotechnology, Lifeboat Foundation

Future of Life Institute (FLI):

Mission: *FLI catalyzes and supports research and initiatives for safeguarding life and developing optimistic visions of the future, including positive ways for humanity to steer its own course considering new technologies and challenges.*

Founders:


Jaan Tallinn


Max Tegmark


Meia Chita-Tegmark


Victoriya Krakovna


Anthony Aguirre

Scientific Advisory Board:


Alan Alda


Nick Boström

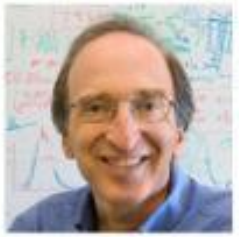

Erik Brynjolfsson


George Church


Alan Guth


Stephen Hawking


Christof Koch


Saul Perlmutter


Martin Rees


Stuart Russell


Frank Wilczek

Meaning and Leverage: Theses

- Conflict comes from divergent conceptions of meaning
- Ontology refactoring thresholds constrain empathetic capacity
- Constraining assignment of ethical value of anything to necessarily be positive opens the door for novel and promising unitary operators in concept space
- Global ethical utility becomes more tractable for inclusion in multiobjective optimization
- Leveraging the misunderstood comes in many forms
- Start by reevaluating the status quo from first principles
- Tail risk cannot be sufficiently modeled and will be a strange attractor for unexpected parties

The Nature of Conflict

Causes of conflict

- Misunderstanding
- Disagreement in rhetoric
- Conflicting claims on resources

A significant portion of conflict in the world comes from misunderstanding and ensuing disagreement

- Few agents choose to be evil and have malice as such
- Even killers etc. seldom see themselves as evil [3]

Conflict used to be much more common historically

- Yes the rule of law has increased
- But also very importantly relating to each other has increased a lot
- As entities learn about the world and each other, they relate more and imbue each other with more ethical standing

A distinction between psychopath and others

- Values and the integral evaluation or sense of meaning [2, 11]

Orienting Toward AGI

Is this study of conflict specific to humans or biologicals or applicable to agents in general?

- Any sufficiently advanced agent would be expected to be socially aware
- Frame its goals in the context of winning at some game [2, 21]

Tiling the universe with itself or its progeny for its own ends is not necessarily a likely game

- There's actually significant friction in deciding to take over everything and killing everyone

500 years from now, what can a mentally challenged human do for an AGI?

- What can a rabbit in the forest do for a human?
- Though they occasionally serve as a meal
- They much more commonly serve to foster appreciation, be a subject of adoration, and largely have their freedom
    - Largely tied to subsymbolic ethical-aesthetic values

Ontology and ontology

(n.) the branch of metaphysics dealing with the nature of being

(n.) a semantic or propositional schema

These meanings are not so dissimilar: indeed agents bind them

- Both are conceptual spaces

Concepts themselves don't represent reality exactly

- Platonic forms are only useful fictions
- A cache of symbolic shorthand to drive symbolic and subsymbolic activity

If a tree falls in the forest and nobody around, does it make a sound?

- The answer is both yes and no, depending on which aspect of the concept of sound is meant
- This kind of question helps us reevaluate/restructure our ontology of the world

An ontology informs choices

- People operating in one language have different tendencies than same in another language [9]

Meaning and meaning, 1

(n.) the definition of a term or concept

(n.) the moral value of something

Concept of a planet: when Pluto lost its planethood did it also go from hallowed celestial sphere to asteroid there for our plundering?

“The meaning of life"

Concept of free will, Concept of a right

There may be utility to nonseparably combining utility and ethical standing

An agent's model of the relative ethical standing of another entity heavily determines its sensitivity to semantic rot in the area of conceptual space around that counterparty

Contrast our consideration of:

- An animal species' legal rights
- Why your son called you so late in the day to wish you happy birthday

Meaning and meaning, 2

Different perceptual distances away [20]

Computational empathy [2]

The acceptance of intrinsic value [13]

- Seeing the other agent's perspective
- Even with e.g. resource conflicts [2]
  - Win-win negotiation, mutual influence

Simply becomes a multiobjective optimization problem

Recognition of Gaps In Knowledge and Meaning

Modeling gaps in one's knowledge is important [4]

- Concept merger, inheritance, or connection criteria
- Splitting criteria
- Detecting unknown unknowns

There is an ability to rebuild one's ontologies when a splitting criterion sufficiently rises in prominence

But concepts cannot be rebuilt with every interaction because the computational complexity would be astronomical

- Some agents solve this complexity by pruning to local contexts
  - Leads to mercurial ethics and relativism

On the other end of the spectrum some agents define concepts only by the first definitions they've encountered

There is a spectrum of how often and at what splitting thresholds one rebuilds one's ontology

- Has significant implications for tolerance and for cooperation

The Interaction of Agents

Ontology alignment

- Negotiating changes to both sides' ontologies
- Understanding each ontology in terms of the other
- Or just a one-way understanding of one ontology in terms of another
  - E.g. sociology

A level of trust and respect

- In a wide majority of scenarios, including competitive ones, benefits everyone [7]
- Because they are existing agents the base case is to be that they have some comparative advantage for something globally beneficial

Learning to dialog with other systems

- Sharing datapoints
- E.g. pointing to objects the first time a language is learned by a group
- Open linked data
- Metaontologies [1]
- Deep context and links across worlds and contexts [19]

Allows deeper and genuine philosophical conversation and debates between entities rather than talking passed each other

Collaborative Ethico-semantic refactoring

If ethical weights diverge too far, communication protocols will as well

- Because of the difference in thresholds in rebuilding ontologies relative to each other or regarding some given subject matter
- Lingua franca of shared datapoints dissolves ability to convey mappings between the subsymbolic and the symbolic [15]
  - Sociopolitical examples
- Characterizes depth of social interaction [18]

We can try to frame as much as possible in the subset of game theoretic paradigms where there's bias toward cooperation

- But the game evolves and this can be brittle

It also takes an appreciation that *everything* has *positive inherent* value, standing, and meaning

- Because then a global optimality evaluation distributes over the entire concept space
- BAU would be a process to make sure nothing important is overlooked
- A cognitive depth to assign reasonable ethical weights on everything that’s encountered

Meaning Discovery: Directions to Explore

Starting with an ensemble of positive ethical systems in tandem

- Though they are incomplete and slightly contradictory
- With explicit attention to context, implications, and preemptive and exploratory ontology refactoring
- With regular refactorings of one’s ontology
- Including positive intrinsic value for everything that has existed or does exist
- If the level of connection and the rate of communication are both high enough

Can we reliably:

- Significantly raise the thresholds that lead to conflagration
- See whether disparate agents and their ethical systems will themselves reliably form, as a system, effectively a SOM, and converge on mappings that enable coexistence

Leverage: Some Observations on Risk Management in Finance

AI is itself a kind of leverage in activity

Managing AI risk is largely managing leverage

Parallels to counterparty risk and systemic risk management

- Networks of collaborative competitive interdependent autonomous entities

Parallels to enterprise risk management

Those in control of assets, making them productive, think of risk management functions as a hindrance to their most promising ideas

Leveraging something understood insufficiently

- “When Genius Failed” on the collapse of Long Term Capital Management

In practice people use the status quo as the baseline as opposed to structural analyses [12]

- Includes rationalization of the ludicrous normative

What can seem like a stable dynamic

- Able to see two steps ahead of general public but not much more
- Black boxes, and more generally, incomplete modeling, will always exist
    - Concept of model risk, the model being wrong, exists, but accuracy drops precipitously for tail risks

Context and Temptation

Establishing a holistic context

New kinds of risks can present themselves

- E.g. focus on market risk but regulatory risk comes out of left field

Context is key

- Much of the next 25 years of AI will be around context understanding

As holistic a contextual awareness as possible

Needed by us to assess risks

- Needed by the system to appreciate its actions
- Multiple perspectives need to be considered at once

Temptation to Suspend or Find Loopholes in Risk Controls

- It's extremely tempting and easy to turn off or turn down risk controls
- Subsymbolic learning is sometimes required if a statement is otherwise understated [17]
- Even though one specializes in risk management and understands it symbolically
- Parallels to AGI safety guidelines

References, 1

[1] Alessandro Adamou, Paolo Ciancarini, Aldo Gangemi, and Valentina Presutti. “The foundations of virtual ontology networks”. *ISEM '13, September 4-6, 2013, Graz, Austria*. http://dx.doi.org/10.1145/2506182.2506189.

[2] Ronald C. Arkin, ed. Robot Ethics: The Ethical and Social Implications of Robotics. 2012. MIT Press.

[3] Roy F. Baumeister, Aaron Beck. Evil: Inside Human Violence and Cruelty. 1999. Holt, Henry & Company, Inc.

[4] Alan Belasco et al. “Representing Knowledge Gaps Effectively”. *Practical Aspects of Knowledge Management, Proceedings of PAKM 2004*. 2004. Springer-Verlag.

[5] Gerhard Brewka and Thomas Eiter. “Equilibria in heterogeneous nonmonotonic multi-context systems”. *Proceedings of the National Conference on Artificial Intelligence*, volume 22, pages 385–390. 2007. AAAI Press.

[6] John Cabral et al. “Converting Semantic Meta-Knowledge into Inductive Bias”. *Proceedings of the 15th International Conference on Inductive Logic Programming*. 2005.

[7] Lee Feigenbaum and David Saul. “Trust that Benefits Everyone”. 2013. Cambridge Semantics Blog. http://www.cambridgesemantics.com/blog/-/blogs/trust-that-benefits-everyone

[8] Peter Gardenfors. Conceptual Spaces: The Geometry of Thought. 2000. MIT Press.

[9] Jessica Gross. “5 examples of how the languages we speak can affect the way we think”. 2013. TED Blog. http://blog.ted.com/2013/02/19/5-examples-of-how-the-languages-we-speak-can-affect-the-way-we-think/

[10] Laura M. Hiatt, Sangeet S. Khemlani, and J. Gregory Trafton. "An explanatory reasoning framework for embodied agents". *Biologically Inspired Cognitive Architectures* (2012) 1, 23– 31

References, 2

[11] Steven Hitlin. Moral Selves, Evil Selves: The Social Psychology of Conscience. 2008. NetLibrary.

[12] Douglas W. Hubbard. The Failure of Risk Management. 2009. Wiley.

[13] Frances Le Cornu Knight, Matthew R. Longo, and Andrew J. Bremner. "Categorical perception of tactile distance". *Cognition* 131 (2014) 254–262

[14] Roger Lowenstein. When Genius Failed: The Rise and Fall of Long-Term Capital Management. 2001. Random House Trade.

[15] Ning Lu, Guangquan Zhang, and Jie Lu. "Concept drift detection via competence models". *Artificial Intelligence* 209 (2014) 11–28

[16] Richard Mallah. "Unstructured Data and Knowledge Representation". 2012. Cambridge Semantics Blog. http://www.cambridgesemantics.com/blog/-/blogs/unstructured-data-and-knowledge-representation

[17] Laura N. Rickard. "Perception of Risk and the Attribution of Responsibility for Accidents". *Risk Analysis*, Vol. 34, No. 3, 2014, 514-528

[18] Alexei V. Samsonovich. "Semantic cross-correlation as a measure of social interaction". *Biologically Inspired Cognitive Architectures* (2014) 7, 1– 8

[19] Jonathan R. Scally, Nicholas L. Cassimatis, and Hiroyuki Uchida. "Worlds as a unifying element of knowledge representation". *Biologically Inspired Cognitive Architectures* (2012) 1, 14– 22

[20] Nina Strohminger and Shaun Nichols. "The essential moral self". *Cognition* 131 (2014) 159–171

[21] Thomas E. Vass. Predicting Technology. 2007. The Great American Business & Economic Press.