From AI to Intelligent Behavior

A Manageable Causal Disconnect

By Leopold B. Willner, PhD October 15, 2019

The consensus is that intelligent behavior must be required of all robots and advanced AI empowered machines, so that humans may feel safe and at ease in their presence. Yet this is not where we are headed with AI and the rise of AGI (artificial general intelligence). Instead, we are racing to put into practice what are mostly context lacking computational tools and methods without consciousness or morality. Such machines include optimizers, recognizers and various other useful computational algorithms. Some are intended to empower advanced autonomous machines and applications whose actions in the real world are not always transparent and yet may be unstoppable. In his seminal book, *Human Compatible*, Cal Berkeley computer science professor Stuart Russell makes it quite clear why this simply will not do. He observes that this shortsighted approach is potentially dangerous to humanity, that is unless we can incorporate 'human preferences' and other controls into the design of all such machines. But can we?

If intelligent behavior is what we require out of our automata, it must by design include ways of incorporating the equivalent of the common human practice of asking 'why and what if' questions within its methods; and do so prior to allowing autonomous AI systems to participate in action in the real world. Yet such forward looking capabilities place broad limitations on self-driving cars and trucks, autonomous: weapon systems, robotic surgery, drug approval and the like. While this may seem straightforward enough for computer science to enable, it is not. Not once we observe that asking 'why and what if' questions is causal and not backward looking and statistical in nature, as described with much clarity and new methods in Judah Pearl and Dana Mackenzie's *Book of WHY*. Thus, this is not a straightforward process to be dealt with by means of, say, statistical correlation and multi-regression tools and the like. Meaning that the path from AI methods to intelligent behavior in an AGI must be tread lightly to avoid the risk of creating out-of-control super intelligent Frankenstein machines leading humankind away from the heart's desire.

Central to all of this is escaping the trap of seeking to make machines into intelligent agents as an end goal in itself, which is both short-sighted and wrong-headed. Instead, inventing ways to achieve intelligent behavior by way of human preferences based autonomous machines, or with the aid of humans-in-the-loop, or in other ways is the logical and safer way ahead. Therein is the promise of a progress that can enable a desirable man-machine duopoly on the path to more intelligent (and safe) behavior on the part of man and machines. This is so because machines can at times be blindsided by the limitations inherent in algorithms, while humans are flawed in other ways due to tribal behavior, a propensity to blindly follow ideologies, their emotions and day to day foibles. Acting together, super intelligent machines focused on human preferences, and humans now learning new lessons from machines and computer science are likely to triangulate on better solutions in the not too distant future.

All the while, what is needed in the very near term, is an 'Intelligent Behavior Platform©' that can support and empower man-machine intelligent behavior applications in healthcare, business, engineering, science, social science, governance and personal affairs. Housed herein, the seminal ideas of professors

J. Pearl, S. Russell and others can lead to the development of valuable new means to light the way to a better, more humanistic future. The modules required in such an IBP may include:

- Preferences
- Causal Modeling
- Context Evaluation
- Cost-Benefit
- Optimization
- Valuation
- Learning
- Feedback & Control

In order to avoid some of the failed experiences of the past, we may wish to recall the experience of an earlier AI, that was based on so-called 'expert systems', heralded as the next coming during the nineteen seventies and eighties at MIT, Stanford, Carnegie Tech and elsewhere. AI based on expert systems was philosophically sound and most promising, but crashed on the rocks of impracticality due to the severe limitations imposed by the laws of large numbers and associated cost. Similarly, while it seems reasonable to suppose that noncarbon based intelligence can in time attain amazing heights, reaching for autonomous intelligent machine behavior seems a bridge too far for now, at least in some key areas. Meanwhile, no amount of analytical discourse is likely to bridge the wide gap between what is big data statistical in nature and what by reason is demanded of valid causal intelligent behavior. The thirty years wasted on expert systems should remind us that what is theoretically possible may not always be practical, as the Concord commercial jet, the personal helicopter and much else have demonstrated.

Without autonomous intelligent machine behavior capability, many AGI concepts are not possible or practical even for highly intelligent machines, and the construction of such automata should be avoided. The hubris of commercial enterprises seeking new markets and greater profits must be moderated by the requirements of humanity for safety - and freedom, privacy and the preservation of human values. Otherwise, we are very likely to create a world that is super-efficient but most unattractive to humankind. Note that in that case the best current human strategy is to shut down all advanced computer science worldwide and render it illegal – with the irony of it all fully exposed!

Instead, where possible and practical, methods that combine human preferences and solutions with those of advanced automata in a symbiotic dance should be sought out. Applications such as these in healthcare, engineering, finance, marketing and science are possible and practical in human-in-the-loop configurations. In other more delicate situations, as in some real time decision making, the joint usage of man and machine results is what is required, with humans mostly remaining in control and purely autonomous machine largely avoided; except in the special cases where causality issues can be fully deconfounded and made manageable. Otherwise, say in a complex surgery that requires why and what-if intelligent behavior and decision making on the part of a surgeon, the autonomous approach is to be avoided.

Failing to do so now or in the distant future, the worst fears of a Bill Gates, Elon Musk, Stephen Hawkins, Stuart Russell and many others could in time be realized, as AGI becomes a menace and not a promising way forward. But such is unlikely if we are wise enough to listen to our wise men, before they are replaced by an even smarter and wiser immortal set of algorithms.

Leopold Willner 10/15/2019 Dual Stream Technology 831-325-5008
Leo@DualStream.Tech DualStream.Tech

REFERENCES

1. Tom Griffiths; *The Artificial Use of Human Beings* in John Brockman; Possible Minds, Penguin Press, pp 125-141, 2019.
2. Stuart Russell, *Human Compatible*; Viking Press 2019
3. Max Tegmark; *Life 3.0, Being Human in the Age of Artificial Intelligence*, Vintage Books 2017.
4. Stuart Russell; *The Purpose Put into the Machine*, in John Brockman; Possible Minds, Penguin Press, pp 20-32, 2019
5. Francesca Rossi; *Building Trust in AI*, Journal of International affairs, Vol 72/No. 1, pp127-133
6. Judea Pearl/Dana Mackenzie; The Book of WHY, *The new science of cause & effect, Basic Books* 2018.
7. Alex Pentland; *The Human Strategy*, J. Brockman-editor, Possible Minds, Penguin Press 2019.
8. *Asilomar AI Principles* (Future of Life Institute, Asilomar 2017), wwws.futureoflife.org/ai-principles/
9. Jim Blascovich and Jeremy Bailenson*; Infinite Reality*, William Morrow of Harper Collins, 2011.
10. Sundar Pichai; *AI at Google: Our Principles* (Google 6/7/2018), blog.google/technology/ai/ai-principles/
11. Trusted AI (IBM 2018), research.IBM.com/artificial-intelligence/trusted-ai/
12. Norbert Weiner; *Human Use of Human Beings, Cybernetics and Society*; Little, Brown Books 1968.
13. Leopold Willner; *Alien Streams*, Research Report 13, Dual Stream Technology 2017. A novel approach to managing AI and AGI at a reduced risk.
14. Pat Langley, *The Cognitive Paradigm*; Advances in Cognitive Systems 1 (2012) 3-13.
15. Peter Voss, *The Third Wave of AI*, Medium.com April 15, 2017.