There are Reasons©

That the Reason may not Always Know

By Leopold B Willner, PhD 8/26/2019

Most of us can agree that voice and face recognition and other such applications are most useful, plus marvel at the fact that AI or artificial intelligence software can beat chess grandmasters and Go gurus! Yet, do these skills lead to, or even toward: creativity, imagination, awareness, feelings and emotions or any real understanding and purposefulness on the part of machines? Most likely not, as an artificial general intelligence (AGI) thus endowed is quite another manifestation that is well beyond defining by means of ordinary computation using big data, mathematical methods and statistical tools and tricks. All the while, ordinary humans, even very small children can do (H)GI quite easily, like grasping context and true meaning.

If this is so, how are we in this new industrial age to move forward with robotics, advanced automation and autonomous systems? Are we simply being led by the profit motives of our corporate giants into an uncertain and possibly undesirable future? That is a challenge that cannot be easily papered over with false hubris or a fear of AGI, both of which can be paralyzing and unhelpful. In the following I add to the growing body of methodology that addresses new ways of improving the prospect of succeeding with such developments.

To begin with it seems a bit of a stretch to assume that there is but one basic form of intelligence in the universe, and that AI and AGI are but varieties on the human theme. Further, that to develop either one, one can confidently mimic the human variety, and add a bit of logic. The reason for this is that any real-world intelligence is a restricted intelligence subject to situational constraints, such as built-in preferences, biases and the like. While logic and mathematics alone may be sufficient to arrive at a purely rational AGI singularity, nothing of the sort is likely to apply to the intelligence of humans here or aliens elsewhere. Indeed, if intelligence exists elsewhere in the universe, which many experts believe is highly probable, it may take different forms in different settings, which are subject to the real world constrains that apply there.

If unity of form among intelligences is a false premise, then AI and AGI may turn out to be quite distinct from human intelligence in significant ways. Thus, we may be wise to not rely too heavily on anthropomorphic models. Meaning that complementarity as in 'Value Alignment' and 'Reinforcement Learning[1,2] and the like – while most valuable in so many applications - may not suffice to always yield the best answers. Instead, 'dual stream triangulation' by different forms of intelligence may provide the sharp contrasts and conflicting views needed to dig out better solutions and outcomes and expose risk in complex problems. That the commercial AI field has so far concentrated on human intelligence modeling may have had more to do with its low hanging fruit than a neglect of such a way forward, via

employing distinct contrasting minds – as with an AI and humans.

None of this suggests that in its ultimate form, as imagined by Max Tegmark in his *Life 3.0*[3] an abstract ideal intelligence may not ultimately emerge within a singularity sometime in the future. But for now, it may be more valuable to allow AI and AGI to develop primarily on first principles so they may act in point and counterpoint to human views, sentiments, sensitivities and values. That is the theme explored in this paper as an adjunct to the more popular and better known top down Bayesian and bottom up big data varieties of AI today.

Such an approach can also do a great deal to assuage the well documented fears of such thinkers as Max Tegmark, Stephen Hawkins and Bill Gates that a nontransparent big-data based AI may lead to a dystopian world for mankind. Instead, as in the view of Anca Dragan[4], getting humans back into the loop can provide better and safer solutions. Daniel Dennett[5] in a similar vein has pointed out that what we want are AI Tools not AI Agents, so that humans may indeed contribute to advanced intelligence while remaining, as needed, fully in charge. Those who may wish to pooh pooh this requirement would do well to consider the in-depth arguments offered by AI pioneer Stuart Russell[6] in support of the idea that an AGI is not only possible but may turn malevolent and be unstoppable[7].

To illustrate that the worries of all these folks is warranted, one need but consider the 2006/2008 near air disasters of three Airbus A330 airplanes in western Australia, and the actual fatal air crashes of two Boeing 737 Max 8 airplanes in Asia the past two years. In each case mostly unknown and inaccessible convoluted AI instructions in automated flight control software forced each of these five planes into dangerous ten degree plus nose dives, mostly without pilot knowledge or direct ability to override. In the auction business that is known as 'fair warning', and in each case Daniel Dennett's guidance should have been applied but was not. Had the human actors been given a greater role, the technocratic hubris that caused these deadly mishaps could and most likely would have been avoided.

Thus, it is fair to suggest that 'there are reasons that the reason may not grasp' in all of AI and AGI, and humans should act to protect themselves from such oversights whether caused directly by an autonomous AI or by its human designers.

With that in mind, AI solutions conjured up by computational methods such as deep neural networks, bounded optimality, enumeration, or gradient descent need to be understood to be what they are: useful tools devoid of conscious understanding which at times may contain a risk to mankind. Thus, effective new ways to truly triangulate between seemingly valid AI results in contrast to 'live' human opinion and sentiment is badly needed, as a stronger and safer approach than AI alone, or one tainted by the fact that it is trained to mimic human behavior out of big data training, Bayesian projections and other methods. Indeed, such a dual approach may become essential if we wish to empower AI tools and robots in ways recommended by Dragan and many others.

When we choose to ascribe intelligence to clever computational tools whether they be Gradient Descent or the machinations of IBM's Big Blue, we are misleading ourselves into assuming that proper goals, objectives, purpose and safety awareness are necessarily built in. They are not, see the Stuart Russell footnote. Indeed, consciousness in humans is there by Darwinian evolution for just these reasons, to pursue opportunities, problems and other purposes in a guided aware state. None of which apply to computer programs, computational procedures, algorithms and the like – unless and until some form of purposeful consciousness can be invented and incorporated into them. A distant event, which may be possible far into the future,

as Tegmark and others have suggested as a possibility.

In the meantime, the powerful new tools found in J. Pearl's '*the new science of cause and effect*'[7] allow us to be optimistic about the rise of (a form of) AGI in the not too distant future. As a consequence, such an AGI should be ever more able to engage and triangulate effectively with HI (human intelligence) to provide safer and more reliable results. It can also be expected to become ever more powerful, rational and unbiased by way of human and autonomous self-training.

One of the sources of confusion about mathematical tools derives from the nature of idealized abstract mathematics, so loved by the author and others, wherein perfection is built in by design in, say, Reimann, Borel and Banach spaces and the like. Here nothing is messy or misunderstood; as mathematicians like Emile Borel and Henri Lebesgue would not have had it any other way. None of which is typically true in the real analog world with its many discontinuities, unpredictable stochastic systems and ever-changing levels of uncertainty. Say as in financial markets and the weather, as well as often irrational and distorted tribal human behavior. Such a situation is lamented by Alex Pentland[9] who notes that algorithms and neural networks can be very powerful; but only iff (as in if and only if) the big data and other tools they employ are guaranteed not to be corrupted or manipulated by humans – say with disinformation. Which is a logical wish, but one that at times is as useful as contemplating truth within a null set!

Indeed, the real world is so very complex and messy that for an AGI to manage it, it must be even more complex than it, and thus, far too complex to be be readily transparent and understood. When human life is at stake, as in self-driving trucks, autonomous weapons, the Boeing 737 Max 8 flight controller, true robotic surgery and much genetic engineering, assuming that what the AGI chooses to do is safe or aligned with human goals and desires is a stretch, and now a source of concern to many, including Elon Musk[10].

The stronger solution set that brings into play two, alien to one another, different forms of intelligence, as say an AI or AGI[11] with a human intelligence, to seek out contrasts and disagreement as well as consent, is a safer and often better approach. Autonomous fly by wire action in the Airbus A320 or Boing 737 Max 8 is made much safer and better when the pilots are in the loop, properly trained and also empowered. Indeed, a village full of 346 dead humans can attest to the fact that this is so, and that the hubris of algorithm designers that led to these tragedies must be counterbalanced by including the possibility of 'live' human action to enable careful formal decision making, as described in the analytical methods of Stanford's Kochenderfer in *Decision Making under Uncertainty*[12]

The future will be based on the safe and effective use of AI and AGI as envisioned by its proponents at Stanford[13], MIT, Google[14], IBM[15] and elsewhere. But this future is at risk unless proper measures are taken to overcome the limitations stated above. With that in mind, a broad-based program to refocus AI in the direction of connecting it to live human intelligence is needed as a way to ensure its future. This can be achieved by means of a reduced dependence on anthropomorphic AI design paradigms along with newly developed dual stream decision under uncertainty capabilities. The complexity to be analyzed and overcome in the development of useful new tools represents a great new challenge. This is a path that should be undertaken in a timely manner with sufficient resources to allow it to succeed. Waiting for the corporate world take up the challenge, while so much profit remains on low hanging branches out of big data and neural networks applications, is unlikely to yield the desired result.

Instead, having an appreciation for the power of the human mind, including its complexity and variety of functions and skills, is a start toward a better understanding of cognition in human intelligence and how this can serve AI. For a penetrating insight into cognition, see the seminal work of Neisser[16]. The way forward to greater automation, robotics, virtual reality along with AI and AGI may in many ways be inevitable, but it can be a risky path to follow if we are careless; for a broad view see Kelly[17].

The conundrum we face today is that we have little choice but to go forward with AI and AGI tools and applications. All the while, we must learn to bring human intelligence and action fully into the equation while we can. We can do so by designing man-machine systems that fully interplay in the way Norbert Weiner[18] intended them to. As with, among others, approaches that exploit innovative dual stream methods[19] and the like. Tools that utilize both AI and human analysis and cognition to reveal when the two differ by way of contrasting views and results. This can act as a much needed 'dead man switch' providing greater safety and efficiency. In conclusion, employing Norbert Weiner's feedback, control and also reflexivity is a way to begin to effectively manage a future cybernetics world.

Leopold Willner PhD 8/26/2019
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. Broadly undermined by education expert Alison Kopnik in *AI Versus Four Year Olds*, in John Brockman; Possible Minds, Penguin Press, pp 219-239, 2019.
3. Max Tegmark; *Life 3.0, Being Human in the Age of Artificial Intelligence*, Vintage Books 2017.
4. Anca Dragon; Putting the human in the AI Equation, John Brockman-Editor, Penguin Press, pp132-141, 2019
5. Daniel Dennett; What Can We Do, John Brockman-Editor, Penguin Press, pp39-53, 2019
6. Stuart Russell; *The Purpose Put into the Machine*, in John Brockman; Possible Minds, Penguin Press, pp 20-32, 2019
7. Francesca Rossi; *Building Trust in AI*, Journal of International affairs, Vol 72/No. 1, pp127-133
8. Judea Pearl; The Book of WHY, *The new science of cause and effect, Basic Books* – Hachette Book Group 2018.
9. Alex Pentland; *The Human Strategy*, John Brockman-editor, Possible Minds, Penguin Press pp193-205 2019.
10. *Asilomar AI Principles* (Future of Life Institute, Asilomar 2017), wwws.futureoflife.org/ai-principles/
11. Preferably an AI or an AGI built on a solid mathematical and psychological foundation of its own, beyond mimicking human behavior and the biases and prejudices that brings in train.
12. Mykel Kochenderfer; *Decision Making Under Uncertainty*, The MIT Press 2015.
13. Jim Blascovich and Jeremy Bailenson*; Infinite Reality*, William Morrow of Harper Collins, 2011.
14. Sundar Pichai; *AI at Google: Our Principles* (Google 6/7/2018), blog.google/technology/ai/ai-principles/
15. Trusted AI (IBM 2018), research.IBM.com/artificial-intelligence/trusted-ai/
16. Ulric Neisser*; Cognition and Reality*, *Principles and Implications of Cognitive Psychology*, W.H. Freeman 1976.
17. Kevin Kelly; *The Inevitable, Understanding the twelve Technological Forces that will shape our future*, Viking 2016.
18. Norbert Weiner; *Human Use of Human Beings, Cybernetics and Society*; Little, Brown Book Group 1968.
19. Leopold Willner; *Alien Streams*, Research Report 13, Dual Stream Technology 2017. A novel approach to managing AI and AGI at a reduced risk.