 DOI: 10.1051/metrology/201709005

Avoiding AI Armageddon with metrologically-oriented psychometrics

Matt Barney[1,*], *William* Fisher[2]

[1]Leaderamp.com, Vacaville, California 95688 USA
[2]LivingCapitalMetrics.com, Sausalito, California 94965 USA

Abstract. With the rapid advancement of powerful artificial intelligence, visionaries such as Stephen Hawking warn that we could be architecting our own extinction. Visionary efforts such as the OpenAI project, and the Ethics and Governance of Artificial Intelligence Fund are key lifeboats to proactively engineer ethics into our technological children, and create other safety strategies to mitigate the likelihood of dangerous AI. Engineering a science of safe AI requires, as a foundational element, an approach to measurement that allows subsequent risk analysis and mitigation methods to be evaluated with meaningful, linear, accurate and precise methods that span a variety of disciplines that contribute to risk variance in organizational, process and team outcomes.

1 Introduction

Specialists in AI are often unaware of organizational and psychological science that is anchored in metrologically oriented psychometrics that are central to mitigating risks of AI harming or killing people. While this social science is powerful, it is largely locked in academic peer-reviewed vaults and proprietary black box applications. Crucially, it also usually lacks the metrological foundations prominent in the physical, biological and agricultural sciences. Ignoring the need for objective measurement in typical high-risk organizational situations has been catastrophic in other domains. For example, Enron, Bernie Maddoff Associates, and MCI/Worldcom collapses lacked the measurement to detect unethical cultures, climates and leadership actions that destroyed these billion-dollar organizations. Similarly, the largest bankruptcy in history, Lehman Brothers, crumbled from lack of visibility to fatal market risks. When organizations fail to anticipate these types of high impact, low probability events, they are fragile [1]. With advanced, dangerous AI we cannot afford to make mistakes [2].

This paper makes the case for a holistic, and interdisciplinary approach to mitigating AI risks based on metrology and metrologically oriented psychometrics. First, we review the transdisciplinary literature on measurement and risk management. Second, we summarize the latest work done jointly by metrologists and psychometricians on the compatibility of measurement across the human sciences. Third, we review the Cue See model of individual, team and organization effectiveness that is anchored in the science of Organizational Psychology, and give examples of computer-adaptive measurement that should be used to screen computer scientists, develop teams, and increase the odds that safe AI organizations and projects realize their goals.

* Corresponding author: matt@leaderamp.com

Article available at http://cfmetrologie.edpsciences.org or https://doi.org/10.1051/metrology/201709005

2 Measurement and risk management

The discipline of Enterprise Risk Management is anchored on the estimation and management of uncertainty. These risks include hazards, such as liability and property damage; financial risks such as liquidity or pricing risk; operational risks such as product failure or customer satisfaction; and strategic risks such as social trends and competitor intellectual property [22]. One principle in risk management is to design organizations and processes to be robust or anti-fragile to organizations. In part, this means that organizations need to be comprised of people and processes that are adaptable to changing environmental circumstances. It also means that organizations need to be able to detect and measure risks, so that prevention and remedy actions are commensurate with their impact and likelihood. No risk management is possible without proper measurement instruments to estimate the uncertainties that may have deleterious consequences on the organization. And some of the biggest possible risks are "intangible" - human resource, senior leader, and culture risks of the sort that destroyed Enron. Consequently, social science measurement is important to detecting and mitigating enterprise risks.

3 Metrologically-oriented psychometrics

Efforts towards applying fundamental measurement theory in psychology date back to Thurstone's work in the 1920s, and made significant advances with the work of Rasch in the 1950s, Luce and Tukey in the 1960s, and others, before and since [4]. Much of this work was produced by Benjamin Wright (1926-2015) and his students. Wright started his career as a physicist, worked with Nobelists Townes and Mullikan, and shifted to psychology in the 1950s. Dissatisfied with the factor analytic methods of the time, Wright sought data and measures that would be "stable in terms that a physicist would accept" [5]. He found that stability in the work of Georg Rasch [6], and helped develop the models, methods, software, and applications putting it into practice, saying years later, "Today there is no methodological reason why social science cannot become as stable, as reproducible, and hence as useful as physics" [4].

Metrology engineers and psychometricians have recently together expressed their agreement with Wright, identifying a basis for a shared language of standards and traceability spanning the natural and social sciences [7,8]. In a special invited address to the international standards laboratories community (NCSLi), Leslie Pendrill, a past chair of the European Association of National Metrology Institutes, said, "The Rasch approach...is not simply a mathematical or statistical approach, but instead [is] a specifically metrological approach to human-based measurement" [9]. Experimentally tested, theoretically explained, and metrologically traceable measures [10] are still more the exception than the rule in psychology and the social sciences, but there are clear trends toward more coherent coordinations and alignment of measures across applications in education, for instance [11-13].

4 Organizational measurement to mitigate AI risk

One approach to organizing the science required to reduce the risks of safe AI teams is an interdisciplinary approach called The Cue See Model [3]. It represents a synthesis of

several scientific disciplines to design and lead organizations to be anti-fragile [1]. It posits that while AI researchers and organizations have specific objectives, they require a variety of assets to realize safety goals. This includes physical, human, and technological assets that must perform together brilliantly to realize goals. But these factors of production are stochastic, and produce emergent outcomes. To achieve those outcomes in a probabilistic system, their performance requirements must be specified at each level of analysis (individual, team and organization) to produce the desired Quality, Cost, Quantity and Cycle Time (QCQC) that ultimately are required for strategic goals to be realized.

In mitigating the risks of evil AI, the primary factor of production and therefore source of variation is creative, ethical problem solving by individuals and teams, which is the focus of study of Organizational Psychology and Organizational Behavior. There is a long, rich study of team problem solving and decision making, with known solutions for mitigating the risks of teams, and leveraging their benefits, through the development of shared mental models, and other forms of team coordination toward shared objectives4. These individual level cognitive, affective, personality and knowledge factors are the key drivers of individual and then team performance that is required in ethical AI efforts. Team communication, mutual performance monitoring, and back-up behavior is similarly required. Ultimately, the organization-level results are an emergent effect of each team, working in their processes, and there are cultural, leadership and governance factors that mitigate the team, process, and individual risks. To manage these complex factors, engineering-worthy measurements must be developed, such as the Rasch family of psychometrics, to detect and mitigate cross-level organizational risks.

Because AI safety innovations are pioneered by individuals, teams and organizations, they are subject to the same sorts of error, and frailty as any other probabilistic human system. Modeling these systems requires good instrumentation to ensure low levels of uncertainty in the inferences we make about individual computer scientists, teams or organizations. While some organizations use metrologically-oriented screening processes for personnel, such as Pilots, with excellent reliability and availability, other high-stakes organizational systems designed to save lives are notoriously unreliable. For example, Johns Hopkins patient safety experts estimate that medical errors are the third biggest cause of death in the USA, causing more than 250,000 unnecessary deaths [14]. And the defect density in Healthcare is an improvement over centuries of healthcare quality improvement efforts, in contrast with AI's rapid and recent advances. Healthcare system risks to safety come from many sources, but the vast majority are from the well-meaning caregivers that lack skills, make mistakes, miscommunicate, or are fatigued.

For organizations building advanced AI, these sources of risk come from individual AI experts, social networks and teams, and from the organizational context itself. Two frameworks have already been proposed to deal with the sort of "Societal Grand Challenge" that strong AI poses, and in this proposal we underscore the need to implement new science and technology to mitigate these risks at the micro, meso and macro levels of analysis, consistent with the Cue See Model of Organizational Effectiveness [3].

4.1 Micro: Individual risks

The United States Department of Labor's (DoL) description of a computer scientist's work and personal attributes identifies many job requirements [15] for an AI expert that may risk their ability to create safe AI. This includes their ability to engineer moral reasoning into strong AI; or invent novel ways to contain the risks of evil AI escaping their laboratories. Many organizations use the DoL as a resource to define work and worker requirements. If that resource is insufficiently broad to account for the worker and job attributes needed for safe AI, that correctable systemic risk should be addressed. While the DoL correctly highlights knowledge and skills that one would expect—mathematics, computers, complex problem solving, and decision making—they do not explicitly include moral reasoning, creativity or persuasion. Because the risks of dangerous AI include ethical and moral considerations, along with innovative safety options and the need to influence non-AI experts to implement solutions, there is a strong case to be made that moral development of AI experts and their colleagues is core to ensuring that Strong AI is as contained as possible.

Fortunately, there is a body of science in psychology around measuring and developing moral reasoning, from psychopathological origins with Anti-Social Personality Disorder, to the highest levels of ethical decision making. Crucially, the best approaches [16-18] employ a measurement approach that is on-par with physical, chemical and biological science requirements for measurement [4-13]. The measurability of a latent trait like moral reasoning is a crucial beginning to being able to cultivate it on mass scales, especially for the ethical decision making skills paramount for individuals working with AI. Advances with computer-adaptive measurement [19] have made it easier to incorporate metrologically aligned psychometrics into everyday work.

4.2 Meso: Team risks

Individual AI skills and ethical decision making may be necessary to increase the odds of creating safe AI, but will still be insufficient to fully mitigate risks. Like the Manhattan project, it will require creative problem solving from a wide range of experts, not limited to AI specialists, on a global scale to solve such a complex problem. For example, it may be that early genetic engineering efforts neglected the need to proactively influence the public for their safety and efficacy, so much that they unintentionally created a large anti-Genetically Modified Organisms (GMOs) movement with backlash and resistance to multiple forms of biotechnology that are sometimes unrelated to real risks with genome-level GMOs.

Unfortunately, the very type of interdisciplinary teams that will be required to address AI risks also notoriously underperform the sum of the contributions of their individuals. They are subject to numerous biases and threats to effective and ethical decision making, such as GroupThink, polarization of decision making, and the Abilene Paradox. For example, when the Nixon Administration officials who funded the 1971 Watergate break-in were questioned about why they approved G. Gordon Liddy's proposal to rob the Democratic headquarters, they explained that his early requests were much worse, and they rejected them all, feeling bad to say no to all his requests, they agreed to the smallest. Eminent persuasion scientist Robert Cialdini notes that Watergate is an example of unethical group decision making that is predictable from his research. Republican leaders

felt the pain that everyone feels by rejecting Liddy's biggest, most immoral requests, and felt that they had to agree to his smallest request because they had rejected him so many times [20]. Conversely, in the right circumstances, team members can provide risk reduction, in the same way as parallel redundant systems do in concurrent engineering. A team with a climate around mutual back-up behavior, can catch AI-related mistakes before they fester, and can challenge flawed moral reasoning that individuals cannot overcome [14].

4.3 Macro: Organizational risks

After Enron, MCI/Worldcom, and Lehman Brother's debacles, the enterprise risk management community has increasingly looked toward interdisciplinary approaches to detecting and mitigating a holistic set of macro-economic, competitor and board/senior leader risks to an organization, where AI is just a special case. But these efforts are still largely ad-hoc, uni-disciplinary and pre-scientific. Enterprise Risk Management (ERM) is dominated by actuarial and financial risk experts who often are unaware of the team, leadership, and organizational cultural risk mitigation science and technology that has developed in psychology and Organizational Behavior. And current methods are grossly insufficient to mitigate the risks of technology that is smarter than humans. For example, with today's technology, estimates suggest that organizations confront at least one serious information security incident per year, and these are often caused indirectly by employees who violate or neglect these policies [21]. Consequently, organization-level risks are threatened by individual employees who make mistakes, breach policy, or make bad decisions.

5 New technologies to mitigate organizational AI risk

Detecting and mitigating these cross-level forms of human risk have historically been difficult because metrologically sophisticated psychometric skills are rare and in high demand with the advent of "Big Data", and the models such as the "Cue See Model", designed to integrate multidisciplinary information about ultimate organizational risks, were cumbersome to implement.

New technologies include three patent-pending applications for mobile platforms incorporating computer-adaptive psychometrics, calibrated coaching, and journaling. These applications support mass-personalization of scarce leadership skills, like the high ethical standards individual Computer Scientists must have to inoculate AI methods with human values; and the ability of highly technical experts to inspire and persuade others to adopt their recommendations. These applications deploy, via iOS and Android, the latest computer-adaptive assessments, and then mass-personalize a psychometrically calibrated form of coaching to help people practice in their daily work. Ideally, this complements expert coaching, to help individuals maximize their potential and performance. An experimental version also blends natural language processing, machine learning and psychometrics to make it even easier for individuals to develop.

These same measurement and coaching methods can be deployed for team, and organization-level interventions, to proactively mitigate the risks of strong AI.

6 Limitations

All novel research efforts have risks, and in this paper we have outlines a variety of hypotheses that we have not explicitly tested. There are a number of limits and risks to the approach we have outlined. First, while we have strong theory from a variety of scientific disciplines to inform our recommended approach, and some of our work is using the latest cloud/mobile technologies to serving live clients, suggesting practical utility. But we have not explicitly tested this approach in the AI organizational domain. It will be crucial to pilot test these approaches, and refine before scaling to a larger audience. Second, without a senior, highly credible AI colleague or visionary such as Elon Musk willing to advocate and sponsor the value of our effort, there will likely be reluctance on the part of AI experts to use our recommendations.

References

1. N. Taleb, *Antifragile: Things that gain from disorder* (Random House, New York, 2012)
2. G. George, J. Howard-Grenville, A. Joshi, L. Tihanyi, Acad. Manag. J. **59**, 1880 (2016)
3. M. Barney, *Leading Value Creation: Organizational Science, Bioinspiration and the Cue See Model* (Palgrave Macmillan, New York, 2013)
4. B. D. Wright, Educ. Meas. Issues and Pract. **16**, 33 (1997)
5. B. D. Wright, Rasch Meas. Trans. **2**, 25 (1988)
6. G. Rasch, *Probabilistic models for some intelligence and attainment tests* (Danmarks Paedogogiske Institut, Copenhagen, 1960)
7. L. Mari, M. Wilson, Measurement **51**, 315 (2014)
8. L. Pendrill, W. P. Fisher, Jr., Measurement **71**, 46 (2015)
9. L. Pendrill, NCSLi Meas.: The J. Meas. Sci. **9**, 22 (2014)
10. W. P. Fisher, Jr., A. J. Stenner, Measurement **92**, 489 (2016)
11. J. Gorin, R. Mislevy R, *Inherent measurement challenges in the next generation science standards for both formative and summative assessment* (ETS, Princeton, New Jersey, 2013)
12. M. Wilson M 2004 *Towards coherence between classroom assessment and accountability* (University of Chicago Press, Chicago, 2004)
13. National Research Council, *Systems for state science assessment* eds M R Wilson and M W Bertenthal (The National Academies Press, Washington, DC, 2006)
14. M. A. Makary, M. Daniel, BMJ **353**, i2139 (2016)
15. National Center for O*NET Development. 15-1111.00. O*NET OnLine. Retrieved June 21, 2017, from https://www.onetonline.org/link/summary/15-1111.00
16. T. Dawson, Internat. J. Behav. Devel. **26**,154 (2002)
17. E. A. Giammarco, Personality Individ. Develop. **88**, 26 (2015)
18. J. J. Haidt, Moral Education **42**, 281 (2013)
19. M. F. Barney, W. P. Fisher, Jr., Ann. Rev. Organiz. Psychol. Organiz. Behav. **3**, 469 (2016)
20. R. Cialdini, *Pre-Suasion: A revolutionary way to influence and persuade* (Simon & Schuster, New York, 2016)
21. L. Myyry, M. Siponen, S. Pahnila, T. Vartiainen, A. Vance, Euro. J. Info. Sys. **18**, 126 (2009)
22. P. Curtis, M. Carey. COSO Risk Assess. in Practice (2012).