Let's Talk AI with Taylor T. Johnson

Taylor T. Johnson[1] and Barbara Steffen[2]

[1] Vanderbilt University, Institute for Software Integrated Systems & Computer Science, USA,
[2] taylor.johnson@vanderbilt.edu
[3] METAFrame Technologies GmbH,
barbara.steffen@metaframe.de

"Formal verification aims to prove whether models satisfy specifications, such as showing a program does what its designer intended. Formal verification is a promising approach that can be used to establish safety, security, and trustworthiness specifications of AI systems. However, to realize the potential societal benefits AI promise, we also need transdisciplinary approaches bridging the gamut from computer science and engineering, the broader sciences, as well as the arts, humanities, social sciences, law, business, and beyond to ensure its development involves all perspectives and voices."

The Interviewee - Taylor T. Johnson

My Personal AI Mission:
To develop formal verification methods to help establish and assure the safe, secure, and trustworthy development and use of AI, especially in the context of safety-critical systems such as autonomous cyber-physical systems (CPS).

My Takes on AI

Artificial Intelligence: I will use the Oxford dictionary definition [1]: "the theory and development of computer systems able to perform tasks that nor-

mally require human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages." An AI system is a computer system that performs such tasks.

Trust: Similarly using an Oxford dictionary definition [2], trust is a "firm belief in the reliability, truth, ability, or strength of someone or something." In the context of AI systems, my view is that it is our ability as users to rely on their results. More specifically to my own research area of verification, it is our ability as designers of AI systems to trust that the AI systems will do what we intend them to do, that is, for them to meet their requirements and specifications.

Explainability: Explainability is complex, as even humans may offer different explanations or reasoning for a decision or outcome than what its true root factors were. Sometimes we may do things without understanding why (or instinctually or with implicit bias), and similarly we may hypothesize why others do things without truly understanding why, and as such, explainability is incredibly difficult. As I am not a psychologist or someone with such expertise, I would define it in the context as a designer of AI systems as our ability to understand what the AI systems are doing based on interrogation and analysis of the systems, predicated on being able to access such systems details, documentation, data, etc.

Essential Elements of Human Capabilities: My view is that AI systems are capable even without replicating human performance on some particular task. However, I would view the essential elements of human capabilities for AI systems to be a true ability for reasoning and generalization across tasks and to new scenarios, not simply meeting or surpassing human performance on some task or a set of tasks. This perhaps would mean artificial general intelligence (AGI) in some form, but I think we are nowhere near such capabilities, as current AI systems do not have understanding of much of anything in the way I believe humans understand, although of course for particular tasks there are impressive recent results with AI. One problematic aspect I find in many recent discussions on AI is that we are tending to anthropomorphize AI, when it is not mimicking human understanding or capability in its current or near-term forms in my view.

The Interview

Barbara *I have the pleasure of speaking today with Professor Taylor Johnson of Vanderbilt University in Nashville. Without any further introductions, I directly hand over to you, Taylor. Could you briefly introduce yourself and your relationship to artificial intelligence?*

Taylor T. Certainly. I'm Taylor Johnson, an associate professor of computer science at Vanderbilt University in Nashville, Tennessee, USA. My research broadly focuses on formal methods and verification, originally for cyber-physical systems. These are modern embedded systems like the computing elements in cars, airplanes, or medical devices. Over the last six or seven years, we've been examining formal verification of machine learning and artificial intelligence systems, specifically components like neural networks [3]. We are particularly interested in how these elements might feature in cyber-physical systems [4]. We explore whether they enable things like autonomous systems in our current reality, such as in self-driving cars and similar systems.

"One of the significant challenges in AI is that we, as system designers, don't fully understand how the systems we're building operate."

Barbara *Interesting, do you have examples of specific challenges you are currently addressing with your AI research?*

Taylor T. Absolutely. As we've seen with recent examples, AI systems don't always operate as intended [5]. We focus on developing new methods, particularly algorithms and software tools, to establish their correctness [6–15]. This is essentially an application of the general formal verification problem in the context of AI systems. We have a model, say a neural network, and a specification of what it should do, and we try to confirm that it does just that. There are numerous challenges in this area, some theoretical, but many practical, like how to define what an AI system should do.

Barbara *In this context, how do you view the role of trust in the adoption of AI, and what measures can you imagine to ensure the ethical use of AI in the future?*

Taylor T. There are many challenges we hope to overcome, particularly in establishing trust. My concept of trust is more related to system designers and engineers. Can these professionals trust the AI systems they're building? The field of verification focuses on assuring the designers that what they're building is correct in some sense. Do the neural networks perform as intended? This is about convincing the designers, less so the end-users. However, if you use a system, say a car, you trust it to operate reliably and not fail mechanically or software-wise. As users, we put our trust in the designers, who are qualified experts, sometimes even licensed, like myself, a professional engineer. We trust these processes for certain systems, like buildings, which we expect not to col-

lapse. Many elements ensure this trust. An analogy often used is Swiss cheese, where we have certification and licensure for the people involved, and they use varying methodologies to establish trust. Ultimately, we have to trust others. One of the significant challenges in AI is that we, as system designers, don't fully understand how the systems we're building operate. There are theoretical gaps on the designer and researcher side. It's complex, but these elements are pieces of the puzzle, filling in the Swiss cheese holes. Other important aspects that I don't personally work on include regulatory and policy matters. Like in our building analogy, in addition to licensing architects and structural engineers, we also have standards and building codes that help establish criteria for proper construction.

Barbara *And in terms of the future capabilities of artificial intelligence, perhaps on a scale of 1 to 10, where 1 refers to artificial intelligence systems dedicated to specific functionalities and purposes, such as ChatGPT. And 10 refers to general artificial intelligence that surpasses human capabilities. Where do you see it going from here, especially in terms of the risks you just mentioned of not knowing exactly what's going on?*

Taylor T. Predicting the future is always challenging. Perhaps I can frame my answer in terms of time frames. Currently, I believe we're at the lower end of that 1 to 10 scale. AI is very proficient at certain specific problems and tasks, so we're probably at a 1 to 3. Over the next decade, I anticipate we might move up to a 5. Beyond that, it's hard to predict. Some of the current systems, like generative AI models and large language models, are transformative technologies with potential impacts similar to the Internet, cell phones, or personal computers. However, they're engineered systems. I find it hard to predict if artificial general intelligence (AGI) will become a reality. I'm not sure if we'll ever reach the 10 on the scale in my lifetime, to align with the theme of this AISoLA conference.

"I believe one of the significant potentials of AI [...] is its ability to impact almost all aspects of life."

Barbara *What are the criteria that need to be met to achieve AGI? How would you measure it?*

Taylor T. Many researchers have considered ways to evaluate whether computers are thinking or their overall capability, which have garnered broader cultural popularity like the imitation game or Turing test. More recently, various large language models (LLMs) have performed well on a variety of tasks and exams that have made headlines in areas we often deem as requiring some level of intelligence, such as test taking. Recently, I attended some research talks that presented interesting ways researchers have defined understanding, for instance in reading comprehension and resulting actions performed. For example, if a robot is given instruction in natural language to pick up an apple, then it does so, it has "understood." There were many instances of course of LLMs clearly not understanding with a variety of criticism around hallucinations and adversarial

prompting, as they are simply generating probable outputs given inputs [16]. My view is current AI systems do not understand in any of the ways humans do. While I am not a neuroscientist nor psychologist, my view is that we are simply anthropomorphizing these engineered systems to try to impart them with capabilities. I do not know all the details, but I do know of some experiments in animal and comparative psychology attempting to understand whether animals even know they exist and evaluate their intelligence, and likely we could debate whether a dog is intelligent or not, or whether a particular dog is more intelligent than another, or whether a particular animal species is more intelligent than another. Perhaps insights from these fields would be directions to consider for how to define AGI or measure progress toward it, as I do think the recent results in things like exam evaluations and other attempts to compare to human or "superhuman" performance—while impressive—are not the way forward. I think anything that will be done for establishing AGI will require interfacing with the real world, in part given the action example, but also for what I view as another essential capability of true intelligence, which is self-preservation (at and beyond both the individual and population/societal levels). These alone are not examples in my opinion for demonstrating intelligence and something more is needed, as AI systems currently do not have any true notion of understanding.

Barbara *And regarding the much-discussed possible future scenarios of artificial intelligence, where do you personally stand on the scale between a more utopian or dystopian view of the future?*

Taylor T. When it comes to future capabilities, while I enjoy dystopian science fiction movies, I believe there are many other societal issues more likely to bring about dystopian scenarios than AI. Transformative technologies over the last century, like the Internet, cell phones, computers, automobiles, and airplanes, have caused problems, but they've also led to improvements. For instance, climate change is an issue related to industrialization and transportation. I view current AI use cases, especially large language models, as tools, much like cell phones, computers, the Internet, automobiles, and airplanes. They have a lot of transformative potential and can lead to efficiency improvements. For example, travel times have drastically reduced due to advancements in transportation. However, these technologies also have the potential to cause problems. I don't foresee a dystopian future, nor a utopian one. That's a broader discussion for socioeconomic considerations and the future of work. I believe AI may enable efficiency, improve our lives, and create new forms of entertainment and art. I don't see killer robots happening. While recent advances in AI, particularly generative models and LLMs, have interesting capabilities, I don't see them leading to a dystopian scenario.

"I don't want to be overly negative about AI. I believe it's a transformative technology, but we have many issues that we need to address before we put it into broad usage."

We have other risks to pay attention to, like the societal issues caused by smartphones and social media. Climate change, resulting from transportation developments, is a bigger concern. I'm optimistic about AI's potential to improve things, but I don't think it'll lead to a utopia where no one needs to work. We derive purpose from our activities, but AI could potentially allow us to work less, which could improve our daily lives. So, I'd say the future lies somewhere between utopia and dystopia, but not fully either.

Barbara *Very interesting and very detailed description of the nuances you see in the further development. We are currently meeting at the AISoLA conference, which looks at artificial intelligence from an international and interdisciplinary perspective. When you reflect on the last few days, are there specific insights from other disciplines that were particularly interesting to you?*

Taylor T. Yes, I believe one of the significant potentials of AI, when compared to the industrial revolution or even developments like the airplane or the automobile, is its ability to impact almost all aspects of life. This is partly due to our ability to communicate globally very quickly now, thanks to the Internet and other telecommunication advances. While these advancements are generally positive, unlike transportation improvements, AI has the potential to directly affect people's lives. One of the key discussions we've had here, which I think is crucial, is the need for interdisciplinary approaches to address this. It's not just about solving engineering problems. We need involvement from people across disciplines, including humanities, social sciences, law, medicine, as well as computer science and engineering, to make these advancements.

"As a broad vision, I believe in developing [...] interdisciplinary fields."

I don't foresee a dystopian scenario happening due to killer robots. However, I do see potential political issues emerging, as AI could lead to consolidation of power amongst companies, governments, or even individuals. This is one of the risks that necessitates an interdisciplinary approach in terms of developing regulatory frameworks, policies, and standards for the engineering and computer science design of these systems. There are unintended use cases and problems that we hadn't considered before some of the recent advancements. This ties into intellectual property law, copyright, and a host of other interesting issues. We need philosophers, artists, historians, writers, computer scientists, and engineers to collaborate and step out of their bubbles, which is one of the great aspects of this AISoLA conference.

For example, in education, AI is being used to generate stories for children to read. This is fascinating, but also potentially problematic if the children learn something strange due to some bias that was either intentional or unintentional in the system. These are some of the significant issues that have arisen in systems like bias in facial recognition systems, as seen in projects like Gender Shades [17]. This has led to issues in policing, credit scoring, loan approvals, and more [5].

AI has the potential to impact all of society, which is different from many other engineered systems we've had. Therefore, it truly requires an interdisciplinary, transdisciplinary, multidisciplinary approach, and we need to step out of our bubbles and engage with others to address these issues. There are many emerging approaches beyond the type of work we have been pursuing at these intersections, such as appropriate systems engineering through documentation and traceability with techniques like model cards and data sheets, as well as auditing and monitoring AI systems [18–21].

Barbara *And now to summarize your wish for the future. From your personal perspective, what is the AI vision that we should address?*

Taylor T. That's a challenging question. As a broad vision, I believe in developing these interdisciplinary fields. As a researcher and scientist, I think that's crucial. Some of this can be achieved through our typical research, but I also think a vision of breaking out of the research community's bubbles would help address some of the potential emerging risks we can see. I don't want to be overly negative about AI. I believe it's a transformative technology, but we have many issues that we need to address before we put it into broad usage. This has shown up in other scenarios, like autonomous vehicles, where some companies recently had their testing permits revoked due to safety concerns. These are great technologies, and I would love to have self-driving cars, but we're not there yet. We need to exercise caution while recognizing that these systems have the potential to significantly change and improve people's lives. In my own research, we have our specific topics, but as a field, I think the vision I would advocate for is to continue developing these interdisciplinary approaches across all fields.

"My view is current AI systems do not understand in any of the ways humans do. [We] are simply anthropomorphizing these engineered systems [...]."

Barbara *Is there anything else you would like to add?*

Taylor T. No, I think we've covered quite a bit.

Barbara *Perfect, then thank you, Taylor, for your insights and your time. Have a great day!*

Taylor T. You're welcome, Barbara. Thank you.

References

1. E. M. Knowles, Ed., The Oxford Dictionary of Phrase and Fable, 2nd ed. Oxford University Press, 2005.
2. A. Stevenson and C. A. Lindberg, Eds., New Oxford American Dictionary, 3rd ed. Oxford University Press, 2010.
3. C. Brix, M. N. Müller, S. Bak, T. T. Johnson, and C. Liu, "First three years of the international verification of neural networks competition (VNN-COMP)," In-

ternational Journal on Software Tools for Technology Transfer, vol. 25, no. 3, pp. 329–339, June 2023.
4. D. M. Lopez, M. Althoff, L. Benet, X. Chen, J. Fan, M. Forets, C. Huang, T. T. Johnson, T. Ladner, W. Li, C. Schilling, and Q. Zhu, "ARCH-COMP22 category report: Artificial intelligence and neural network control systems (AINNCS) for continuous and hybrid systems plants," in Proceedings of 9th International Workshop on Applied Verification of Continuous and Hybrid Systems (ARCH22), ser. EPiC Series in Computing, G. Frehse, M. Althoff, E. Schoitsch, and J. Guiochet, Eds., vol. 90. EasyChair, 2022, pp. 142–184.
5. I. D. Raji, I. E. Kumar, A. Horowitz, and A. Selbst, "The fallacy of AI functionality," in Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, ser. FAccT '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 959–972.
6. W. Xiang, H.-D. Tran, and T. T. Johnson, "Output reachable set estimation and verification for multilayer neural networks," IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 11, pp. 5777–5783, 2018.
7. H.-D. Tran, D. Manzanas Lopez, P. Musau, X. Yang, L. V. Nguyen, W. Xiang, and T. T. Johnson, "Star-based reachability analysis of deep neural networks," in Formal Methods – The Next 30 Years, M. H. ter Beek, A. McIver, and J. N. Oliveira, Eds. Springer, 2019, pp. 670–686.
8. H.-D. Tran, F. Cai, M. L. Diego, P. Musau, T. T. Johnson, and X. Koutsoukos, "Safety verification of cyber-physical systems with reinforcement learning control," ACM Trans. Embed. Comput. Syst., vol. 18, no. 5s, October 2019.
9. H.-D. Tran, S. Bak, W. Xiang, and T. T. Johnson, "Verification of deep convolutional neural networks using imagestars," in Computer Aided Verification, S. K. Lahiri and C. Wang, Eds. Springer, 2020, pp. 18–42.
10. H.-D. Tran, X. Yang, D. Manzanas Lopez, P. Musau, L. V. Nguyen, W. Xiang, S. Bak, and T. T. Johnson, "NNV: The neural network verification tool for deep neural networks and learning-enabled cyber-physical systems," in Computer Aided Verification, S. K. Lahiri and C. Wang, Eds. Springer, 2020, pp. 3–17.
11. S. Bak, H.-D. Tran, K. Hobbs, and T. T. Johnson, "Improved geometric path enumeration for verifying ReLU neural networks," in Computer Aided Verification, S. K. Lahiri and C. Wang, Eds. Springer, 2020, pp. 66–96.
12. H.-D. Tran, N. Pal, P. Musau, D. M. Lopez, N. Hamilton, X. Yang, S. Bak, and T. T. Johnson, "Robustness verification of semantic segmentation neural networks using relaxed reachability," in Computer Aided Verification, A. Silva and K. R. M. Leino, Eds. Springer, 2021, pp. 263–286.
13. X. Yang, T. T. Johnson, H.-D. Tran, T. Yamaguchi, B. Hoxha, and D. Prokhorov, "Reachability analysis of deep ReLU neural networks using facet-vertex incidence," in Proceedings of the 24th International Conference on Hybrid Systems: Computation and Control, ser. HSCC '21. New York, NY, USA: Association for Computing Machinery, 2021.
14. X. Yang, T. Yamaguchi, H.-D. Tran, B. Hoxha, T. T. Johnson, and D. Prokhorov, "Neural network repair with reachability analysis," in Formal Modeling and Analysis of Timed Systems, S. Bogomolov and D. Parker, Eds. Springer, 2022, pp. 221–236.
15. D. M. Lopez, S. W. Choi, H.-D. Tran, and T. T. Johnson, "NNV 2.0: The neural network verification tool," in Computer Aided Verification, C. Enea and A. Lal, Eds. Springer, 2023, pp. 397–412.
16. E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in Proceedings of the 2021

ACM Conference on Fairness, Accountability, and Transparency, ser. FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 610–623.
17. J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in Proceedings of the 1st Conference on Fairness, Accountability and Transparency, ser. Proceedings of Machine Learning Research, S. A. Friedler and C. Wilson, Eds., vol. 81. PMLR, February 2018, pp. 77–91.
18. M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in Proceedings of the Conference on Fairness, Accountability, and Transparency, ser. FAT* '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 220–229.
19. I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, "Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing," in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, ser. FAT* '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 33–44.
20. T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. III, and K. Crawford, "Datasheets for datasets," Commun. ACM, vol. 64, no. 12, p. 86–92, November 2021.
21. S. Costanza-Chock, I. D. Raji, and J. Buolamwini, "Who audits the auditors? recommendations from a field scan of the algorithmic auditing ecosystem," in Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, ser. FAccT '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 1571–1583.