Models are Central to AI Assurance

Robin Bloomfield
City, University of London
and Adelard, part of NCC Group
London, UK
r.e.bloomfield@city.ac.uk | robin.bloomfield@nccgroup.com

John Rushby
Computer Science Laboratory
SRI International
Menlo Park CA, USA
Rushby@csl.sri.com

***Abstract*—All interactive systems need a model of their world that they can use to calculate effective behavior. For assurance, the model needs to be accurate but, in autonomous vehicles and many other AI applications, the model is built by a perception system based on machine learning and the *dependability perspective* maintains that its accuracy cannot be assured. We outline this perspective and methods for providing assurance using guards and defense in depth, and we also outline predictive processing as a possible way to construct assured models. We then discuss LLMs, which typically lack explicit models of the world, and suggest possible mitigations for their correspondingly unpredictable behavior. Finally, we consider models in AGI.**

***Index Terms*—assurance, predictive models, autonomy, LLMs**

I. Introduction: the Dependability Perspective

This paper outlines two of the themes developed in our recent technical report on Assurance for AI systems [1]; please note that we deliberately reuse some of its text. A major theme of the report is to distinguish what we call the *dependability* and the *trustworthiness* perspectives on AI assurance. The dependability perspective derives from successful traditional methods for construction and assurance of critical systems. Its basis is that those who develop, assure, and evaluate such systems must have near-complete understanding of how the given system works, what are its hazards, how these are eliminated or mitigated, and how we can be sure all this is implemented correctly. The evidence and structured arguments that justify *indefeasible confidence* in the claims documenting this understanding constitute an *assurance case* [2].

AI and, particularly, Machine Learning (ML) components do not conform to these requirements because they are developed heuristically: their implementation (e.g., weights in a large neural net) is constructed experimentally by "training" on a set of examples. The hope is that if the system works correctly on the training examples, then it will work correctly on all similar examples. The dependability perspective asserts there is no way to provide strong assurance for this: the implementation is opaque and not amenable to analysis, while testing cannot probe more than a fraction of the real example space in feasible time. The contrary trustworthiness perspective does claim some assurance is possible for the behavior of AI and ML components that have been developed, analyzed, tested, augmented, or restricted in various ways (e.g., formal methods can verify low-dimension neural networks [3] but not those of the scale and complexity of interest here, which also lack formal specifications).

Both perspectives have merit and there is a continuum or spectrum between them and potential value in their combination. Our report and this paper focus on points toward the dependability end of the spectrum; we invite others to consider additional points along the spectrum.

II. Models

Models are another theme of our report; they are ubiquitous in science and engineering and are used for many different purposes (e.g., to understand the world or describe a design) [4]; our focus is on the *predictive models* that guide a system's behavior. Any system that interacts with the world must have such a model of relevant aspects of the world [5]. Its model allows the system to predict the evolution of the world and hence to select actions that will advance its goals. For example, the model for a simple control system will be a set of differential equations that describes the interaction of its controlled plant and its environment. Typically, these simple models will be used during design but will be represented only indirectly, as control laws, in the actual system. More complex systems will add state machines to their models, while cyberphysical systems will use integrated formalisms such as timed and hybrid automata, and sometimes high-fidelity design models will be employed, as in "digital twins" [6]. Some elements of these models will be represented directly in the system and some may even be determined or adjusted at runtime, as in adaptive and model-predictive control. And sometimes the entire model will largely be constructed at runtime and be represented explicitly within the system, as in autonomous systems.

In an autonomous system, and we will take self-driving cars as an example, the model will represent the local road layout and the locations and relevant attributes of other road users, pedestrians, and obstructions. Unlike a traditional system that determines the state of its model with sensors that can be "read" directly, a self-driving car must build its model with a *perception system* that uses AI and ML to interpret sensors such as cameras and lidars. The lower levels of perception are model-free: the camera does not "know" that it is looking at a road and its ML-based object classifier has only an implicit model (because its training was composed entirely of road images). The predictive model is built by higher levels of perception that are deliberately programmed to interpret the classified objects as meaningful entities on and around a road.

Partially funded by SRI International.

A separate *action system* uses the model to calculate and execute driving decisions that are safe and will advance a (human-defined) goal. This action system may also use AI and ML but it is possible to build a *guard* or monitor that checks its output for safety. The guard can be built using traditional methods and can be highly assured from the dependability perspective, but it will depend on the model (because it essentially simulates proposed actions against the model and checks their safety).

This approach does not provide full assurance: for that we must either provide assurance for the model, or base the guard on some other, assured model. We will tackle the latter approach first.

III. Assured Guards

One idea is to provide very simple guards, rather like the Automatic Emergency Braking (AEB) that will be required on new passenger cars and light trucks in the USA by September 2029. This is basically a classical control system, with a simple sensor (e.g., radar) and control laws derived from a model of vehicle dynamics. It is feasible to provide strong assurance for AEB within a limited *Operational Design Domain* (ODD), which we will call a μODD (pronounced micro-ODD) [7], such as "forward collisions on highways at less than 80 mph."

An argument against this approach is that it can trigger unnecessarily, either as a false alarm (perhaps being outside the intended μODD, or because the primary self-driving system was planning some other avoidance maneuver), is harsh (modest speed reduction might have done the job if triggered earlier), does nothing for other hazardous μODDs, and may even be hazardous itself (e.g., precipitate a rear-end collision).

One mitigation for unnecessary and harsh interventions is to provide *defense in depth* by introducing a *safety system* between the primary and emergency systems. The safety system uses AI and ML to build an alternative (e.g., simpler) model than the primary system and can override it (if implemented as part of a dual-process architecture, to be discussed in Section IV) or be fused with it. For example, the primary system of a self-driving car might see an object ahead and classify it as a cardboard box and be prepared to drive over it; as it comes closer, the emergency system will see it as an unclassified obstruction and slam the brakes on; but earlier, a safety system could have seen it as an unclassified but possibly hazardous obstruction, and caused the action system to change lanes to avoid it "just in case."

Because it uses ML, the model in the safety system cannot be assured from the dependability perspective, but it is reasonable to suppose (and to attempt to ensure) that it is *diverse* from the primary model, and likely to fail somewhat independently.[1] Furthermore, the claim that the safety system needs to support is not that it ensures safety (that is assured by the emergency system), but that it reduces demands on the emergency system. Defense in depth poses challenging problems in design: generally all layers of the system must be developed as a whole to ensure that differences among them do not cause unnecessary loss of availability, while nevertheless preserving system safety arguments based on diversity.

An enhanced variant on defense in depth provides multiple emergency systems for different μODDs (e.g., for forward collisions on highways vs. in city traffic). There may be an unacceptable risk of false alarms if all are active, so we can add a selector to switch among them, or to inhibit those considered inappropriate. The selector will likely require ML to detect the appropriate μODDs, but this seems a simple function requiring only an elementary model where the trustworthiness perspective may provide adequate assurance [9].

[1]The topic of assurance through diversity is large and somewhat contentious. There is little doubt that architectures employing diverse components are generally more reliable than single threads. The difficulty is in demonstrating that diversity provides benefit in any *particular* case, and in estimating *how much* benefit it provides [8].

IV. Assured Models

Guards with elementary models can provide only crude protection and the model of an intermediate safety system may be difficult to fuse with the primary model; a better solution is to find a way to assure the primary model.

Conventional perception systems work "bottom up": one or more deep neural nets take sensor data (e.g., images from cameras or point clouds from lidars) and deliver interpretations (e.g., lists of detected objects) that are further processed and fused to produce the world model. A fundamental problem with this approach is that it works "backwards" from effect (images) to cause (objects), which is inherently difficult and therefore prone to failure, as exhibited by *adversarial examples* [10]. Another problem is that this approach prioritizes fleeting sensor data above the world model, which is the repository of much accumulated information.

An alternative approach, and the way human perception is believed to work [11], reasons "forwards" using a model to *predict* sensor data (or low-level interpretations thereof) and then applies a form of Bayesian inference to optimize the model in a way that minimizes future *prediction error* (this is rather like a Kalman Filter, but for complex data). Notice that predictions provide sensors with the model that is lacking in bottom-up interpretation (e.g., the model tells the lane detector where to start looking for the markings), while prediction errors provide continuous feedback on accuracy of the world model. Furthermore, minimization of prediction error provides a principled way to perform fusion over diverse lower-level sensor and perception functions.

In humans, this mechanism for perception is known as *Predictive Processing* (PP) [12] and it is believed to be coupled with a *dual-process* architecture [13]. The lower-level process, known as "System 1" [14], performs rapid unconscious perception using PP so long as prediction errors are fairly small, indicating the world is evolving as expected. A large prediction error (e.g., the unpredicted appearance of an obstruction) is called a *surprise* and the higher-level "System 2" process intervenes to resolve it using more deliberative cognition or an alternative model. We recommend this architecture for autonomous systems [15].

V. Models in LLMs and Similar Tools

We have described how world models are essential to autonomous systems and sketched how assurance can be developed for systems whose primary models are constructed using AI and ML. Similar architectures and methods can be applied to other systems that are developed for specific purposes and use AI and ML: the specific purposes (and corresponding assumptions and context) allow us to determine the hazards and hence the safety claims for which the system must be assured and for which it must build models that its guards and backups can use to monitor the action system.

Matters become more difficult when we consider general-purpose AI building blocks such as Large Language Models (LLMs) and Diffusion Models (for images), partly because we do not know the context of their possible deployments, and hence cannot anticipate specific hazards and their mitigations. In our report, we describe how general-purpose systems can be weakly guarded for very general properties, such as ethics, law, and reputation, and internally constrained by additional training against a "constitution" [16].

Here, we focus on the fundamental problem of general-purpose AI, which is that LLMs, for example, are *model-free*. Effective cooperation and communication among humans is based on the parties having a shared context or world model and some awareness of each other's version of that model. For example, if I am your driving instructor, I need a model of your model of the UK Highway Code. An LLM has none of this: its utterances align with the world models of any specific context purely by statistical association. Hence, the utterances of LLMs are unconnected with the collaborative context; they are what philosophers call *bullshit* [17] and frequent failures (misleadingly called "hallucinations") are to be expected.

On the other hand, LLMs are popular because their performance goes far beyond that suggested by their training goal to "predict the next or missing 'token' in a string of text": they have emergent behavior that delivers more value than this. Similarly, although there are no explicit models providing context for this behavior, *implicit models* possibly emerge from statistical associations in LLM training data, and this might partially account for their surprising performance. It would also explain their flaws (implicit models are unpredictable) and, further, indicate that these flaws are irregular, inevitable [18], and unfixable—unless assurable world models can be incorporated within explicit guards or within LLMs themselves.

One approach is to examine the implicit model and purported reasoning that the LLM does employ, which it can be engineered to disclose as an *explanation* [19]. In some applications (e.g., generation of formal text such as designs or code), the explanation may be sufficient to guide an external simulator or verifier to assure the output (e.g., prove a generated program is correct). If the verification fails, then a counterexample can sometimes be constructed and returned to the LLM with instructions to try again (and again). Variations invite the LLM to critique its own output, or to follow a step-by-step plan that reflects the user's own model.

A variation subjects the LLM output to scrutiny by a diverse AI system. Typically this will use symbolic methods such as deduction over a (human-generated) model for the intended application domain, possibly augmented with the ability to lookup trusted Web sites. Some applications of LLMs are already incorporating these techniques as "plugins" (e.g., Microsoft's Bing is reported use over 100 plugins [20]).

These approaches use external models to guard LLMs. Other approaches constrain the LLM to apply a model that is explicitly provided. One such method exploits the large *context window* (i.e., input) allowed by some recent LLMs and provides a prompt with hundreds of training examples prior to the real query. This is called *in-context learning* [21]; previously, such "fine tuning" required access to the LLM's training environment and adjusted the weights in its neural net. Related to this are applications that provide a substantial input and then ask the LLM to do something with it (e.g., summarize it, or identify its topic). These approaches constrain the LLM to operate on or within the context provided, so there is little opportunity to generate or insert inappropriate content extracted from its training corpus. Due to its model-free nature, the LLM may still misinterpret the input and do a *wrong* thing, but this should also be minimized as a large input can explicitly (via extensive prompts) or implicitly (via a block of text) convey the intended context or model.

Notice that sometimes it may be desirable to adjust the data presented to an AI or ML system as a way to manipulate its implicit model. For example, if racial bias is recognized as a hazard, then it might be mitigated by removing race from the data presented to the ML in training and operation. A weakness in this approach is that the ML may discover a proxy for race (e.g., zip code) among the data that it does see, so a better alternative may be to mask this characteristic in training by assigning race randomly. An alternative could be to repeat queries under different assumptions—such as with race or gender assigned differently—and compare decisions. These methods can be seen as computational approximations to Rawls' "veil of ignorance" [22].

All these approaches provide compensation for the model-free behavior of LLMs and other general-purpose AI systems. The utility and safety of these systems would be greatly improved if methods were developed to incorporate suitable models into their construction and behavior. In the interim, the external methods sketched here do have the merit of diversity and can provide limited assurance on that basis.

VI. Artificial (Fairly) General Intelligence

In previous sections we have considered assurance for specific systems and for general-purpose tools based on AI and ML where the concern is that faulty behavior may lead to harm. However, for projected developments of current AI and ML systems and the potential emergence of *Artificial General Intelligence* (AGI—a hypothesized capability where AI performance exceeds that of humans on many tasks), the concern is not just faulty behavior but the social impact of new capabilities. In particular, an AGI system capable of setting its

own goals might pursue objectives contrary to human interests. Notice that although the terminology is seldom employed in discussion of these topics, these are nonetheless dependability failures: an AI system with potentially contrary goals is a hazard that should be recognized and it should be furnished with safety requirements to mitigate the danger, together with assurance that they do so, and are implemented effectively. However, current practice in the field frames the assurance problem for AGI as ensuring that its goals *align* with those of human society—and this needs to be maintained even though AGI may fall into the hands of bad actors, criminals, and adversaries.

In our opinion, current AI and ML technology will not lead to AGI, precisely because it is model-free: a true AGI needs an accurate predictive model of its world. What we consider of more urgent concern is Artificial *Fairly General* (or *Fairly Good*) Intelligence (AFGI) that is a modest projection of the technology already available. These are AI systems that are good enough and cheap enough to displace (possibly superior) human services. For example, well-researched journalism already finds it hard to compete with LLM-generated press releases and propaganda that masquerade as news. Furthermore, by repeatedly circulating misinformation, falsehoods, and mediocrity, these systems impair our own ability to discriminate truth, novelty, and insight, and they contaminate the training data consumed by the next generation of LLMs, possibly leading to *model collapse* [23].

In our report, we develop these and other dystopian possibilities, and also the hazard of AI system becoming conscious. However, we are unable to suggest mitigations other than voluntary or regulated vigilance by those involved in their development. But we do stress that provision of world models, and AI perception systems that construct assured models, will be the central concern.

VII. Conclusions

We identified a spectrum of approaches to assurance for AI systems, ranging from the traditional dependability perspective, which holds that AI and ML lack the predictability required for assurance and must instead be externally guarded, to the trustworthiness perspective, which believes these systems can sometimes be adequately assured directly.

We focused on methods toward the dependability end of the spectrum and argued that the central issue is assurance for the predictive models of the system's local environment that are used to drive its behavior. Because the accuracy of models constructed by an AI perception system cannot be guaranteed, the dependability perspective favors guards that use simpler, but assured models. In autonomous systems, these guards may be safe but disruptive, so we recommended a near-term strategy of defense in depth using architectures with diverse assured guards. Longer term, development of assured perception, possibly based on predictive processing, is the important challenge.

For LLMs, the concern is not that they misperceive the world and build inaccurate models, but that they have no world model at all and are therefore completely unpredictable. In the near term, diverse external guards that do employ a model can provide some mitigation, but the longer term challenge is to develop LLMs that infer explicit and checkable world models from their training, or that have suitable models imposed upon them during their development. Finally, we glanced at AGI and suggested that dystopian future prospects should not distract from the near-term hazards of fairly good AI that is fragile.

Acknowledgments

We are grateful for helpful advice on our report received from Cliff Jones and Brian Randell of the University of Newcastle upon Tyne and from Phil Koopman of CMU, and for constructive comments on this paper by the anonymous reviewers.

References

[1] R. Bloomfield and J. Rushby, "Assurance of AI systems from a dependability perspective," arXiv:2407.13948, 2024.
[2] ——, "Confidence in Assurance 2.0 Cases," in *The Practice of Formal Methods: Essays in Honour of Cliff Jones, Part I*, Springer LNCS Vol. 14780: Sep. 2024, pp. 1–23.
[3] G. Katz *et al.*, "Reluplex: An efficient SMT solver for verifying deep neural networks," in *Computer-Aided Verification, CAV '2017, Part I*, Springer LNCS vol. 10426: Jul. 2017, pp. 97–117.
[4] E. A. Lee, *Plato and the Nerd: The Creative Partnership of Humans and Technology*. MIT Press, 2017.
[5] R. C. Conant and W. R. Ashby, "Every good regulator of a system must be a model of that system," *International Journal of Systems Science*, vol. 1, no. 2, pp. 89–97, 1970.
[6] *Foundational Research Gaps and Future Directions for Digital Twins*. National Academies Press, 2023.
[7] P. Koopman, B. Osyk, and J. Weast, "Autonomous vehicles meet the physical world: RSS, variability, uncertainty, and proving safety," in SafeComp, Springer LNCS vol. 11698: Sep. 2019, pp. 245–253.
[8] B. Littlewood and D. R. Miller, "Conceptual modeling of coincident failures in multiversion software," *IEEE Transactions on Software Engineering*, vol. 15, no. 12, pp. 1596–1614, Dec. 1989.
[9] H. Torfah *et al.*, "Learning monitor ensembles for operational design domains," in *Runtime Verification (RV 2023)*, Springer LNCS vol. 14245: Oct. 2023, pp. 271–290.
[10] C. Szegedy *et al.*, "Intriguing properties of neural networks," arXiv:1312.6199, 2013.
[11] M. Bennett, *A Brief History of Intelligence: Evolution, AI, and the Five Breakthroughs that Made Our Brains*. Mariner Books, 2023.
[12] W. Wiese and T. K. Metzinger, "Vanilla PP for philosophers: A primer on Predictive Processing," in *Philosophy and Predictive Processing*, T. K. Metzinger and W. Wiese, Eds. Frankfurt am Main: MIND Group, 2017.
[13] K. Frankish, "Dual-process and dual-system theories of reasoning," *Philosophy Compass*, vol. 5, no. 10, pp. 914–926, 2010.
[14] D. Kahneman, *Thinking, Fast and Slow*. Farrar, Straus, Giroux, 2011.
[15] S. Jha, J. Rushby, and N. Shankar, "Model-centered assurance for autonomous systems," in *Computer Safety, Reliability, and Security (*SafeComp*)*, Springer LNCS vol. 12234: Sep. 2020, pp. 228–243.
[16] Y. Bai *et al.*, "Constitutional AI: Harmlessness from AI feedback," arXiv:2212.08073, Dec. 2022.
[17] M. T. Hicks, J. Humphries, and J. Slater, "ChatGPT is bullshit," *Ethics and Information Technology*, vol. 26, no. 2, p. 38, 2024.
[18] Z. Xu, S. Jain, and M. Kankanhalli, "Hallucination is inevitable: An innate limitation of Large Language Models," arXiv:2401.11817, Jan. 2024.
[19] R. Dwivedi *et al.*, "Explainable AI (XAI): Core ideas, techniques, and solutions," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–33, 2023.
[20] S. Lazar, "Frontier AI ethics: Anticipating and evaluating the societal impacts of generative agents," arXiv:2404.06750, Apr. 2024.
[21] T. Brown *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.
[22] J. Rawls, *A Theory of Justice.* Belknap/Harvard University Press, 1971.
[23] I. Shumailov *et al.*, "AI models collapse when trained on recursively generated data," *Nature*, vol. 631, no. 8022, pp. 755–759, Jul. 2024.