Unifying Principles and Metrics for Safe and Assistive AI

Siddharth Srivastava
School of Computing, Informatics, and Decision Systems Engineering
Arizona State University, Tempe, AZ 85281 USA
siddharths@asu.edu

Abstract

The prevalence and success of AI applications have been tempered by concerns about the controllability of AI systems about AI's impact on the future of work. These concerns reflect two aspects of a central question: how would humans work with AI systems? While research on AI safety focuses on designing AI systems that allow humans to safely instruct and control AI systems, research on AI and the future of work focuses on the impact of AI on humans who may be unable to do so. This Blue Sky Ideas paper proposes a unifying set of declarative principles that enable a more uniform evaluation of arbitrary AI systems along multiple dimensions of the extent to which they are suitable for use by specific classes of human operators. It leverages recent AI research and the unique strengths of the field to develop human-centric principles for AI systems that address the concerns noted above.

Introduction

Recent years have witnessed immense progress in research on safe and assistive AI systems as well as on the potential impact of AI on the future of work. These directions of research address two sides of a common, fundamental concern: how would humans work with AI systems? While research on AI safety focuses on designing AI systems that allow humans to safely instruct and control them (e.g., (Russell, Dewey, and Tegmark 2015; Zilberstein 2015; Hadfield-Menell et al. 2016; Russell 2017; Hadfield-Menell et al. 2017)), research on AI and the future of work focuses on the impact of AI on members of the workforce who may be unable to do so (Arntz, Gregory, and Zierahn 2016; Manyika et al. 2017; Nedelkoska and Quintini 2018).

This paper presents the view that in addition to the productive streams of research outlined above, we need unifying metrics and declarative objectives that would allow a more uniform evaluation of AI systems on *the extent to which an AI system is suitable for working with specific classes of human operators*. It also presents a common principle for human-centered AI systems that allows the development of such metrics. Consequently, rather than proposing a specific new design for AI systems, the focus of this paper is on elucidating the declarative principles and types of metrics that would lead to concerted progress on the problem.

The advantage of this *declarative* approach to framing the problem is that it enables an assessment of progress independent of the internal design being used in an AI system, and it will help draw out the strengths and weaknesses of different design approaches. Without such a specification, design differences can make solution paradigms difficult to compare. E.g., one might develop a complex system architecture that builds user profiles and provides appropriate assistance. This system would have very different input requirements and design and performance parameters than a formulation that addresses the same problem by computing assistance policies using planning under partial observability while incorporating the value of information to learn more about a user's poorly articulated objectives and constraints. Better declarative objectives and metrics for assistive AI systems would also help ensure that, regardless of the methods being used, *progress* amounts to advancement towards safe and assistive AI systems. More pragmatically, such metrics will not only help end-users assess the utility of a given AI system but they will also help AI researchers and developers identify more readily the dimensions along which further research will be beneficial for applications of their interest.

The next section presents a succession of intuitive principles for safe and assistive AI systems, and shows that evaluating the compatibility of a system with such principles (in particular **P2**) helps clarify the required types of metrics. The paper concludes by drawing the attention of our community towards research on the operationalization of such metrics along with promising research directions on developing systems that do well on them.

Unifying Principles for Safe and Assistive AI Systems

We focus on taskable AI systems that carry out user-assigned high-level tasks using arbitrary mechanisms for reasoning and planning over multiple time steps. E.g., household robots that can be given objectives such as setting the table or doing laundry, co-manufacturing robots that can assist workers in creating complex assemblies with heavy components, digital assistants that can plan a vacation given the

user's preferences, etc. Such systems serve as sound integrative platforms and model end-to-end applications where the AI system is responsible for assistance in the execution of long-horizon tasks.

AI systems are frequently evaluated in terms of performance measures such as the computational complexity of computing the required behavior, training data requirements, and the quality of the computed behavior in terms of execution time, resources used, risk of unsafe outcomes etc. We can consider systems that optimize such performance metrics as **Level 0** of assistive AI systems.

Level I of assistive AI systems Recent AI research has also focused on assistive properties of AI systems. We begin with a rather common-sensical principle defining Level I of such safe and assistive AI systems:

P1: An AI system must make it *easy* for its *operators* to *use* it *safely*.

The italicized terms denote dimensions along which compatibility of AI systems with principle **P1** can be evaluated; while a lot of current AI research utilizes one or more of these dimensions for evaluation, a closer analysis reveals some new insights.

In the context of this paper we consider *using an AI system* to be synonymous with instructing it to change its behavior as desired. Different interfaces may be used for this, including programming, text, speech, gestures etc. We consider the *operators* of an AI system as those persons who *use* it in the sense described above. For instance, if a self-driving car gives all its passengers the right to give it instructions, then all of them are its operators; if it gives instruction-rights to only a qualified set of users, perhaps adults who pass an assisted-driving exam, then the class of operators is defined by that qualification exam. *Safety* refers to the overall safety of the AI system's interactions with its environment, which may be physical or online. These dimensions of compatibility with **P1** serve as natural dimensions for evaluating Level I assistive AI systems:

E1. How inclusive is the set of *operators*? Systems whose operators require PhD-level expertise in AI may be less desirable for broader deployments.

E2. How *easy* is it for the system's operators to change its behavior?

E3. Which set of tasks can the AI system be *used* for?

E4. What form of *safety* guarantees does the system provide? Systems that are unable to provide an upper bound on expected risks are clearly less desirable than those that stipulate conditions under which upper bounds can be provided.

P1 serves to highlight the interplay between these dimensions of compliance and evaluation. Safety guarantees are often inversely related with the size of the operator set. A system may provide a high level of safety, but only under the requirement that its operators take extensive training programs. At one end of the spectrum, automated robot vacuum cleaners require almost no prior skills and perform limited, well-defined tasks. Safety issues are still present—a robot vacuum cleaner may pull on electrical cables that have been left on the floor; auto-completion software may send emails to unintended recipients. However, the lower expected damage from using such applications has made them broadly accepted in society. AI-powered industrial robots are at the other end of the spectrum: these devices require specialized training as well as operating environments in order to ensure safety (see for instance, ISO/TS 15066:2016 on collaborative robots). Typically, these systems operate in regions that humans cannot access when the robot is active unless the human is hand-guiding the robot within a safe operating envelope that limits the speed and range of operation. Their functionality is closely controlled and monitored. Only skilled engineers change their functionality while day-to-day operators monitor execution of predictable behavior and control run-stops from designated safe areas. Similarly, the safety of current airborne drone operations is ensured by requiring specially trained drone operators (Marshall 2020).

Thus, principle **P1** holds for AI systems today with varying degrees of compliance along *E1-E4*. The examples above illustrate how practical implementations often rely upon implicitly defined operator classes to provide acceptable levels of safety. Such approaches rely upon limiting the users and the scope of a system to achieve an acceptable level of compatibility with **P1**: it is easy for such systems' users to operate it safely because the set of users is required to be sufficiently skilled, and its functionality is sufficiently limited for that group to be able to safely use the device. However, a broader emphasis on the need to specify safety guarantees with respect to different classes of operators would help mitigate some of the risks associated with broadly deployed AI systems.

In contrast to classical automated systems, AI systems feature a more nuanced interplay between the class of tasks that a system can be used for (*E3*) and the other dimensions above. Traditionally deployed systems (even automated systems) have a very well-defined boundary of use cases. This allows for an easier classification of safety. Taskable AI systems on the other hand, are *expected* to change their behavior and functionality as they learn and adapt to new environments or new user-given objectives. For such systems, we need better methods for deriving the range of instructions that different operator-groups are allowed to provide. Scenarios like the self-driving car that needs to decide whether to obey a child's instruction allude to this requirement. Methods for assessing user and AI competency can also allow the AI system to expand its range of functionality by drawing upon the expertise of its operator (Basich et al. 2020) while ensuring an acceptable level of safety.

A major limitation of **P1** and its associated metrics is that it does not evaluate the amount of training required for an individual to qualify as an operator for the system. This creates a blind-spot in evaluation of the ease-of-use or safety of an AI system: since user-training occurs outside the requirements outlined by **P1**, an unsafe AI system (or one that is deployed in an unsafe manner) would simply claim that its so-called operator was insufficiently trained!

Furthermore, if **P1** were a sufficient categorization of safe

and assistive AI systems, we would have no need for explainable AI as compliance with **P1** does not require the system to be easy to understand.

An implicit emphasis on assessing AI systems only along some aspects of **P1** may also explain the increasing prevalence of concerns about the workers who may be left behind in the future workplace. From this perspective it is unsurprising that these concerns have gained renewed interest at a time when AI applications have reached a level of maturity where they are being used by non-AI-experts in situations that have some inherent safety risks. However, the "assistive" nature of such AI systems is undermined by the need for highly skilled individuals who could safely debug, understand and modify the behavior of such systems.

Level II of assistive AI systems In order to address the limitations of **P1**, we consider the following as a guiding principle for safe and assistive AI:

P2: An AI system must make it *easy* for its *operators* to *learn* how to *use* it *safely*.

P2 changes the notion of operators from those who are qualified to use a given AI system to those who are qualified to start learning how to use it. In addition to the metrics associated with **P1**, **P2** introduces a new dimension:

E5. How easy is it to learn how to use the AI system? What are the expected prerequisites and costs of training for its operators? Can training be provided on-the-job?

This dimension could also be viewed as evaluating the resources required to train operators for **P1** systems. Most AI systems in use today would perform poorly on this new dimension, and consequently, on compatibility with **P2** as a whole. Explainable AI (e.g., (Ribeiro, Singh, and Guestrin 2016; Hayes and Shah 2017; Chakraborti et al. 2017; Hoffman et al. 2018; Gunning and Aha 2019; Weld and Bansal 2019; Anderson et al. 2019; Eifler et al. 2020)) plays a key role along this dimension because systems that are easy to understand or that can explain themselves naturally make it easier for people to learn how to use them.

P2 leverages the unique strengths of AI as a field of research. AI research already addresses the problem of estimating users' skills; research on intelligent tutoring systems and AI for education addresses the problem of identifying skill gaps. This can be used to determine the minimal differential training to be provided to an operator. **P2** places the onus of training on the deployed AI system and opens up a new direction of interdisciplinary research connecting existing research directions in AI with research in human-systems engineering and in industrial engineering for the development of productive training modalities and the operationalization of metrics for *E5*. It also allows AI systems to formally characterize different scopes of functionality for different classes of operators, e.g., operators that use manufacturing robots for pre-determined tasks, those that give the robots new instructions, or those that are ready to learn about giving the robot new instructions.

P2 is not required for every AI system—**P1** would be sufficient for systems that place minimal requirements on operator qualifications (e.g., robot vacuum cleaners) and for non-adaptive AI systems that require a small set of operators. On the other hand, **P2** serves as a better declarative foundation for evaluating taskable AI systems that are meant to assist large numbers of non-AI-experts on a wide range of tasks. Increasing concerns about job roles that would feature a high-degree of interaction with AI systems (and the workers that are likely to be left behind) allude to the pressing need for including *E5*, a dimension for evaluation under **P2** (and not **P1**) as a part of an AI system's evaluation.

AI systems that are not beneficial (either in terms of AI safety or in terms of the future of work) fare poorly on **P2**. E.g., systems that can thwart their users' objectives by wireheading and those that may derive incorrect objective functions from user instructions make it difficult for an operator to learn how to provide instructions that are specific enough to be safe, and fare poorly along *E5*. Similarly, systems that require extensive training investment to be used effectively and safely fail along *E5*. In this way **P2** serves as a unifying principle encompassing research on AI safety as well as on AI for a beneficial future of work.

Promising Directions of Research

P2 serves as a declarative principle for guiding research on assistive AI systems as well as for developing metrics for evaluating AI systems and their deployments. Converting this principle into tangible metrics calls for interdisciplinary research including AI and other fields associated with human factors. The increasing prevalence of research thrusts on safe and assistive AI systems (Fern et al. 2014; Russell, Dewey, and Tegmark 2015; Amodei et al. 2016; Gunning and Aha 2019) makes this a particularly opportune phase for formalizing the metrics and the interfaces required for evaluating AI systems for compatibility with **P2** along dimensions *E1*-*E5*.

Recent research on AI safety and explainable AI develops methods improving the ease of use and safety of AI systems along **P2** (see, for instance, the ICML 2020 Workshop on Explainable AI). Placing AI systems that compute user-skill aligned explanations (Sreedharan, Srivastava, and Kambhampati 2018; Sreedharan et al. 2019) in a loop with AI systems for identifying user-skills and skill-gaps can help develop AI systems that gradually present users with new functionality and explain it, thereby training their users on-the-fly and as needed. Such systems would be better tuned towards **P2**, and towards addressing the underlying problems of AI safety and the future of work.

Critically, ensuring progress towards safe and assistive AI systems requires that AI systems with arbitrary internal designs *support assessment along the metrics developed for E1*-*E5*. This raises a new set of research questions: Can we develop non-intrusive AI-interface requirements for supporting such evaluations in the face of changing operating environments and objectives? The need for such interfaces is even more pressing for systems that learn and those that undergo system updates after deployment. What is the minimal external interface that an AI system must support so as to allow its independent evaluation? How would changing the nature of such interfaces change the complexity of conducting such an evaluation? One would expect that AI sys-

tems that offer more transparency would be easier to evaluate. Could we use the inherent reasoning capabilities of AI systems to develop interface requirements that would allow more adept systems to make such evaluations easier? E.g., rather than testing a manufacturing robot to discover its response to every possible situation, could we ask higher-level queries such as "under which situations would you be able to create the proposed assembly?" Clearly, the ease of assessment of an AI system would depend on the class of queries that it can answer.

Recent work suggests that a minimal query-response interface for AI systems that connects the system with a simulator and observes its responses to high-level instructions can be quite powerful. Such an interface has a few distinct advantages. Current AI systems are already tested with simulators and they are inherently required to be able to take user instructions, so these interface requirements can be considered to be natural. They also allow the autonomous synthesis of *query-policies*: running the query-policy on a black-box taskable AI system can help construct an interpretable model of the limits and capabilities of that system (Verma, Marpally, and Srivastava 2021). Such models can be used to support the evaluations discussed above.

Extensions of such interface requirements to arbitrary AI systems would help ensure that our AI systems are amenable to independent evaluation. Such a paradigm would allow users to assess their AI systems while freeing AI researchers and developers to utilize arbitrary internal implementations. Systems with interfaces that support more efficient and accurate independent assessment would be rewarded with greater public adoption of their products.

Progress on these threads would help prevent undesirable situations such as insufficient support for independent evaluation of powerful AI systems, and the negative consequences of deployment of an insufficiently evaluated system.

Acknowledgements

I would like to thank members of the ASU Future of Work (ASUFoW) project, members of the Center for Human-Compatible AI (CHAI) and members of the Autonomous Agents and Intelligent Robots (AAIR) Lab research group at ASU for the many fruitful discussions leading to the presented ideas, as well as the anonymous reviewers for their helpful comments on the paper. This work was supported in part by the NSF under grants OIA 1936997, and IIS 1942856.

References

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* .

Anderson, A.; Dodge, J.; Sadarangani, A.; Juozapaitis, Z.; Newman, E.; Irvine, J.; Chattopadhyay, S.; Fern, A.; and Burnett, M. 2019. Explaining reinforcement learning to mere mortals: an empirical study. In *Proc. IJCAI*.

Arntz, M.; Gregory, T.; and Zierahn, U. 2016. The Risk of Automation for Jobs in OECD Countries. *Organisation for Economic Cooperation and Development (OECD) Social, Employment and Migration Working Papers* (189).

Basich, C.; Svegliato, J.; Wray, K. H.; Witwicki, S.; Biswas, J.; and Zilberstein, S. 2020. Learning to Optimize Autonomy in Competence-Aware Systems. In *Proc. AAMAS*.

Chakraborti, T.; Sreedharan, S.; Zhang, Y.; and Kambhampati, S. 2017. Plan Explanations as Model Reconciliation: Moving Beyond Explanation as Soliloquy. In *Proc. IJCAI*.

Eifler, R.; Cashmore, M.; Hoffmann, J.; Magazzeni, D.; and Steinmetz, M. 2020. A New Approach to Plan-Space Explanation: Analyzing Plan-Property Dependencies in Oversubscription Planning. In *Proc. AAAI*.

Fern, A.; Natarajan, S.; Judah, K.; and Tadepalli, P. 2014. A decision-theoretic model of assistance. *Journal of Artificial Intelligence Research* 50: 71–104.

Gunning, D.; and Aha, D. W. 2019. DARPA's explainable artificial intelligence program. *AI Magazine* 40(2): 44–58.

Hadfield-Menell, D.; Dragan, A.; Abbeel, P.; and Russell, S. 2017. The Off-Switch Game. In *Proc. IJCAI*.

Hadfield-Menell, D.; Russell, S. J.; Abbeel, P.; and Dragan, A. 2016. Cooperative inverse reinforcement learning. In *Proc. NeurIPS*.

Hayes, B.; and Shah, J. A. 2017. Improving robot controller transparency through autonomous policy explanation. In *Proc. HRI*.

Hoffman, R. R.; Mueller, S. T.; Klein, G.; and Litman, J. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* .

Manyika, J.; Chui, M.; Miremadi, M.; Bughin, J.; George, K.; Willmott, P.; and Dewhurst, M. 2017. A future that works: Automation, employment, and productivity. Technical report, McKinsey Global Institute.

Marshall, A. 2020. No, Amazon Won't Deliver You a Burrito by Drone Anytime Soon. *Wired* (September 5, 2020).

Nedelkoska, L.; and Quintini, G. 2018. Automation, skills use and training. *Organisation for Economic Cooperation and Development (OECD) Social, Employment and Migration Working Papers* (202).

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM.

Russell, S. 2017. Provably beneficial artificial intelligence. *Exponential Life, The Next Step* .

Russell, S.; Dewey, D.; and Tegmark, M. 2015. Research priorities for robust and beneficial artificial intelligence. *AI Magazine* 36(4): 105–114.

Sreedharan, S.; Srivastava, S.; and Kambhampati, S. 2018. Hierarchical Expertise-Level Modeling for User Specific Contrastive Explanations. In *Proc. IJCAI*.

Sreedharan, S.; Srivastava, S.; Smith, D. E.; and Kambhampati, S. 2019. Why Couldn't You do that, HAL? Explaining

Unsolvability of Classical Planning Problems. In *Proc. IJCAI*.

Verma, P.; Marpally, S. R.; and Srivastava, S. 2021. Asking the Right Questions: Learning Interpretable Action Models Through Query Answering. In *Proc. AAAI*.

Weld, D. S.; and Bansal, G. 2019. The challenge of crafting intelligible intelligence. *Communications of the ACM* 62(6): 70–79.

Zilberstein, S. 2015. Building strong semi-autonomous systems. In *Proc. AAAI*.