You should not control what you do not understand: the risks of controllability in AI

Gabriel Diniz Junqueira Barbosa[1][0000−0003−3929−2736] and
Simone Diniz Junqueira Barbosa[1][0000−0002−0044−503X]

PUC-Rio, Rua Marques de Sao Vicente, 225, Gavea, Rio de Janeiro, RJ, Brazil
gabrieldjb@gmail.com, simone@inf.puc-rio.br

Abstract. In this paper, we posit that giving users control over an artificial intelligence (AI) model may be dangerous without their proper understanding of how the model works. Traditionally, AI research has been more concerned with improving accuracy rates than putting humans in the loop, *i.e.*, with user interactivity. However, as AI tools become more widespread, high-quality user interfaces and interaction design become essential to the consumer's adoption of such tools. As developers seek to give users more influence over AI models, we argue this urge should be tempered by improving users' understanding of the models' behavior.

Keywords: Controllable AI · Explainable AI · Risks of Controllability · Human-AI Interaction.

1 Introduction

Human-Computer Interaction (HCI) is becoming increasingly concerned with how users interact with artificial intelligence (AI) models [2]. Usual considerations of HCI apply: How does a user interact with AI models? Can they understand how these models' decision-making processes work? Do they trust AI-based tools? Should they trust them? These are just a few concerns within the HCI community about how humans and AI may interact.

As AI tools become more widespread in commercial settings, industry is starting to notice how poor user experience – in regard to human-AI interaction – can act as a barrier. Users may have trust issues with tools that exclude them from the decision-making process, as well as very high expectations regarding their performance [3]. It might be tempting for industry to yield some control over these models to appease users, but this urge may lead to graver consequences.

Control without understanding is dangerous. Users that engage with systems they do not understand are more prone to errors [9, 5]. Depending on the AI model's responsibilities, the negative consequences of these errors may end up being more severe [6].

2 Transparency & Understanding

Users often do not understand how artificial intelligence works. This results in a mostly exploratory use of AI-based systems. In certain contexts of use, this is not a problem. However, as tasks executed by users and AI become more important, exploratory use starts to become a greater problem. An individual testing out controls becomes more prone to errors, with potentially harmful results [9].

Learnability is an essential aspect of human-computer interaction [9]. Learning often takes place in controlled environments, *e.g.*, through tutorials or reversible actions. This process allows the user to try different commands without fear of negative consequences. However, AI's behavior is either unpredictable or too complex for humans to predict. This makes it more difficult for users to understand model behavior through trial and error [7].

The behavior of machine-learning models also depends on the data being input to the model. In real usage scenarios, users of a model do not have prior knowledge about the data used to generate it, nor do they know what kinds of input data the model can process effectively. If their learning process is limited to trial and error, it becomes more difficult for the users to anticipate the possible outcomes in these novel scenarios.

Some systems are too complex for trial and error. A user may have to spend an enormous amount of time testing possibilities until he/she understands how the AI model works [7]. These models need to be more explainable, so as to make it easier for users to grasp the basics of model behavior. These explanations usually involve some degree of simplification. It is important not to simplify too much, however, otherwise the explanation may not be precise enough to explain specific model behaviors [11].

Explainable models must also be transparent, so as to allow the user to evaluate how they are operating and thus assess which outcomes are more trustworthy. In this context, transparency may also help in user learning [1].

Explaining models to users is also context dependent. Different models and contexts of use may require different explanations. So do different users. A mathematician does not require the same level of simplification as a child. It then becomes paramount for interaction designers to conduct user research, and understand how stakeholders use these tools, so as to create explainable models more adjusted to the users' profiles and circumstances [11].

Users ought to have some understanding of the model's behavior prior to being given control over it. Exploratory behavior may end up being harmful [6], and controlled learning environments can be inefficient in helping users understand model behavior [7]. Proper explanation requires designers who understand stakeholders' needs and can create different ways to explain model behavior [11].

As the users start to understand the model, they become less likely to err when given control over it. Understanding possible outcomes allows the user to avoid making risky changes, therefore promoting a conservative ("safe") approach to their interactions with the model [1].

3 Controllability

As defined by Roy et al. [8], controllability is the amount of control a user has over an AI model. Traditionally, users would not have much control over model behavior. Once models have been configured or trained, they would make decisions autonomously. However, as users increasingly engage with AI-based technology, this autonomous behavior has been met with suspicion [12, 3].

Users do not appreciate being left out of decisions. Even if they do not want to affect the outcomes, they want to be afforded the opportunity to do so. Shneiderman, in his 1997 discussion with Mae, argues that users seek a feeling of mastery and responsibility, and not the sense that they were not helpful to the process [10].

To ensure higher user satisfaction, developers may be tempted to allow users to control some aspects of AI models. As mentioned above, doing so before the user has proper understanding of model behavior may be dangerous.

There are different ways to give the user control. Developers may give them control over the outcomes, or control over the models themselves [8]. The latter is more complex, as it requires better explanations and understanding of model behavior.

In machine-learning models, users configure the model training by tuning its hyperparameters. This allows them to input their own preferences and create a model that is compatible to their preferences and experience [4]. However, once these models are trained, changing them would require retraining. Moreover, the users of a trained model may not have access to information about how the model was trained, and therefore would be unaware of limitations and biases.

All of these control scenarios may result in errors if the user does not sufficiently understand the model behavior. Through different explanations, it is possible to increase users' understanding of the model, therefore allowing them to exert some control over it [11].

4 Discussion

In this paper, we argued that, although controllability in AI is generally considered desirable, giving users control over AI models without ensuring they have a proper understanding of the models' behavior may lead to dire outcomes. Depending on the situation in which these AI models are implemented, these outcomes may be catastrophic [6]. It is therefore important to develop ways to make models transparent and explain their behavior to users.

Once users understand better how these models work, they will be less prone to making mistakes. They may then be given control, resulting in less undesirable outcomes. Different models may allow for different control methods, with some being more permissive than others [8].

In the end, no one solution will fit all situations. AI models are quite different from one another, and each requires specific methods of explanation and control. Users are also very diverse, so it is important to understand for whom these models and explanations are being designed.

Users want more control over AI models and outcomes in their tools. However, if the models are not properly explained and users do not understand how they work, this control may end up being catastrophic.

References

1. Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P.N., Inkpen, K., Teevan, J., Kikin-Gil, R., Horvitz, E.: Guidelines for Human-AI Interaction. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 3:1–3:13. CHI '19, ACM, New York, NY, USA (2019). https://doi.org/10.1145/3290605.3300233
2. Inkpen, K., Chancellor, S., De Choudhury, M., Veale, M., Baumer, E.P.S.: Where is the Human?: Bridging the Gap Between AI and HCI. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. pp. W09:1–W09:9. CHI EA '19, ACM, New York, NY, USA (2019). https://doi.org/10.1145/3290607.3299002
3. Kocielnik, R., Amershi, S., Bennett, P.N.: Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI Systems. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 411:1–411:14. CHI '19, ACM, New York, NY, USA (2019). https://doi.org/10.1145/3290605.3300641
4. Linden, G., Hanks, S., Lesh, N.: Interactive Assessment of User Preference Models: The Automated Travel Assistant. In: Jameson, A., Paris, C., Tasso, C. (eds.) User Modeling. pp. 67–78. International Centre for Mechanical Sciences, Springer Vienna (1997)
5. Norman, D.: The Design of Everyday Things: Revised and Expanded Edition. Basic Books (Nov 2013)
6. O'Neil, C.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown, New York, 1 edition edn. (Sep 2016)
7. Phelan, C., Hullman, J., Kay, M., Resnick, P.: Some Prior(s) Experience Necessary: Templates for Getting Started With Bayesian Analysis. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 479:1–479:12. CHI '19, ACM, New York, NY, USA (2019). https://doi.org/10.1145/3290605.3300709
8. Roy, Q., Zhang, F., Vogel, D.: Automation Accuracy Is Good, but High Controllability May Be Better. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 520:1–520:8. CHI '19, ACM, New York, NY, USA (2019). https://doi.org/10.1145/3290605.3300750
9. Sharp, H., Preece, J., Rogers, Y.: Interaction Design: Beyond HumanComputer Interaction. John Wiley & Sons, Indianapolis, IN, edio: 5th edn. (2019)
10. Shneiderman, B., Maes, P.: Direct manipulation vs. interface agents. interactions **4**(6), 42–61 (Nov 1997). https://doi.org/10.1145/267505.267514
11. Wang, D., Yang, Q., Abdul, A., Lim, B.Y.: Designing Theory-Driven User-Centric Explainable AI. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 601:1–601:15. CHI '19, ACM, New York, NY, USA (2019). https://doi.org/10.1145/3290605.3300831
12. Zhou, J., Li, Z., Hu, H., Yu, K., Chen, F., Li, Z., Wang, Y.: Effects of Influence on User Trust in Predictive Decision Making. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. pp. LBW2812:1–LBW2812:6. CHI EA '19, ACM, New York, NY, USA (2019). https://doi.org/10.1145/3290607.3312962