Human-AI joint task performance: Learning from uncertainty in autonomous driving systems

DOI:
10.1016/j.infoandorg.2024.100502

Document Version
Final published version

Information and Organization 34 (2024) 100502

Contents lists available at ScienceDirect

Information and Organization

journal homepage: www.elsevier.com/locate/infoandorg

Human-AI joint task performance: Learning from uncertainty in autonomous driving systems

Panos Constantinides [a,*], Eric Monteiro [b], Lars Mathiassen [c]

[a] Alliance Manchester Business School, University of Manchester, UK
[b] Norwegian University of Science and Technology, Norway
[c] Georgia State University, USA

ARTICLE INFO

Keywords:
AI systems
Human-AI joint task performance
Uncertainty
Learning
Tesla
Autonomous driving systems

ABSTRACT

High uncertainty tasks such as making a medical diagnosis, judging a criminal justice case and driving in a big city have a very low margin for error because of the potentially devastating consequences for human lives. In this paper, we focus on how humans learn from uncertainty while performing a high uncertainty task with AI systems. We analyze Tesla's autonomous driving systems (ADS), a type of AI system, drawing on crash investigation reports, published reports on formal simulation tests and YouTube recordings of informal simulation tests by amateur drivers. Our empirical analysis provides insights into how varied levels of uncertainty tolerance have implications for how humans learn from uncertainty in real-time and over time to jointly perform the driving task with Tesla's ADS. Our core contribution is a theoretical model that explains human-AI joint task performance. Specifically, we show that, the interdependencies between different modes of AI use including *uncontrolled automation*, *limited automation*, *expanded automation*, and *controlled automation* are dynamically shaped through humans' learning from uncertainty. We discuss how humans move between these modes of AI use by increasing, reducing, or reinforcing their uncertainty tolerance. We conclude by discussing implications for the design of AI systems, policy into delegation in joint task performance, as well as the use of data to improve learning from uncertainty.

1. Introduction

In recent years, we have witnessed an increase in the scope of tasks performed by AI systems (Berente, Gu, Recker, & Santhanam, 2021; Jain, Padmanabhan, Pavlou, & Raghu, 2021). AI systems can "perform cognitive functions that we associate with human minds, such as perceiving, reasoning, learning, interacting with the environment, problem solving, decision-making, and even demonstrating creativity" (Rai, Constantinides, & Sarker, 2019, p. iii). AI systems are sociotechnical in that they involve humans, who design, train and validate algorithmic technologies based on user-generated and synthetic data.[1] Today's AI systems learn to perform tasks

* Corresponding author.
E-mail addresses: panos.constantinides@manchester.ac.uk (P. Constantinides), eric.monteiro@ntnu.co (E. Monteiro), lars.mathiassen@ceprin.org (L. Mathiassen).

[1] Here we acknowledge that, for example, data used to train an AI system for autonomous driving depends on both data generated by human drivers and synthetic data generated in computer simulations. Most AI systems depend on both such data. We also acknowledge that humans who design and train algorithmic technologies may differ from those that validate these technologies. Most AI systems depend on inputs from machine learning experts, professional users (e.g. professional drivers) and lay users (e.g. amateur drivers).

https://doi.org/10.1016/j.infoandorg.2024.100502
Received 22 November 2021; Received in revised form 22 December 2023; Accepted 15 January 2024
Available online 30 January 2024

autonomously and can outperform even expert humans at an increasingly long list of tasks across occupational categories (Eloundou, Manning, Mishkin, & Rock, 2023; Felten, Raj, & Seamans, 2023). This increased AI system performance presents humans with a novel choice: perform tasks on their own or delegate (partly or fully) these tasks to AI systems.

Previous research has shown that humans tend to choose humans over AI systems – a phenomenon called "algorithm aversion" (Burton, Stein, & Jensen, 2020; Castelo, Bos, & Lehmann, 2019; Dietvorst, Simmons, & Massey, 2015; Filiz et al., 2023; Reich, Kaju, & Maglio, 2022). Such aversion takes place in tasks that are perceived by humans to be uncertain, i.e., these tasks are open to interpretation and based on personal opinion or intuition, whereas low uncertainty tasks have quantifiable and measurable parameters (Burton et al., 2020; Castelo et al., 2019). For example, recommending dating partners is perceived to exhibit high uncertainty, whereas providing financial advice on mortgage loans can be statistically measured and risks can be potentially mitigated (cf. Mousavi & Gigerenzer, 2014). In addition, research has found that humans are more prone to avoid using AI systems for tasks perceived to have more serious consequences if performed with errors (e.g., evaluation of MRI scans and criminal case files) (Filiz et al., 2023). In fact, early studies on algorithm aversion have found that when AI systems make errors, humans tend to avoid using them altogether (Dietvorst et al., 2015). The assumption is that AI systems cannot learn from errors in the same way that humans can (Reich et al., 2022).

Humans in the experiments studied are asked whether they would use an AI system given a description of the system and its performance parameters vis-à-vis human performance (Castelo et al., 2019; Filiz et al., 2023; Reich et al., 2022). This is very much dependent on their *perception* of joint task performance, not on their actual joint task performance. So, even though there is acknowledgement that humans can learn about how AI systems perform, something that allows humans to adjust their perceptions of the uncertainty involved in performing different tasks, these studies have not examined the impact of learning on human-AI joint task performance. Even more recent studies that examined the performance of AI systems against standardized academic aptitude tests as a measure of task exposure to automation (Eloundou et al., 2023; Felten et al., 2023), cannot provide an understanding of human-AI joint task performance in situ. We thus ask, *how does learning from uncertainty impact human-AI joint task performance?*

To answer this research question, we draw on recent empirical research that examines the actual use of AI systems in organizations to theorize how humans learn to jointly perform tasks with AI systems in high uncertainty tasks (Hartmann & Wenzelburger, 2021; Jussupow, Spohrer, Heinzl, & Gawlitza, 2021; Lebovitz, Lifshitz-Assaf, & Levina, 2022). In this effort, we are concerned with the ways in which humans learn to use AI systems in real-time and over time. We define learning as a change in actionable decision-making based on experience that accumulates through "trial and error … and the selection and retention of past behaviors" (Gavetti & Levinthal, 2000. p. 113). In other words, learning is based on experience, whereby humans reinforce behaviors through exploration and exploitation of different options (Gottlieb & Oudeyer, 2018; Walker, Luque, Le Pelley, & Beesley, 2019). Research in this tradition suggests that humans learn based on their individual *uncertainty tolerance*[2] (Griffin & Grote, 2020). For example, humans who are intolerant of uncertainty are likely to avoid or seek to reduce uncertain situations; whereas humans who are tolerant of uncertainty tend to be more explorative of uncertain situations (Cohen, McClure, & Yu, 2007). These differences in uncertainty tolerance are key to the learning dynamics in joint human-AI task performance in ways we analyze in detail.

We conduct an analysis of Tesla's autonomous driving systems (ADS), a type of AI system, drawing on (a) crash investigation reports on these systems provided by the US National Highway Traffic Safety Administration (NHTSA) and the US National Transportation Safety Board (NTSB), (b) published reports on formal simulation tests of Tesla's ADS, and (c) YouTube recordings of informal simulation tests by amateur drivers activating Tesla's ADS in major cities. Our empirical analysis provides insights into how humans' varied levels of uncertainty tolerance have implications for how they learn from uncertainty in real-time and over time to jointly perform the driving task with Tesla's ADS.

Our core contribution is a theoretical model that explains human-AI joint task performance. Our model supports that "augmentation cannot be neatly separated from automation… these …are interdependent across time and space" (Raisch & Krakowski, 2021: 193). Specifically, we show that, the interdependencies between different modes of AI use including *uncontrolled automation*, *limited automation*, *expanded automation*, and *controlled automation* are dynamically shaped through humans' learning from uncertainty. Humans learn from their own errors and the errors made by AI systems. Understanding how those errors manifest in real-time and over-time and with what consequences helps to theorize subsequent learning for human-AI joint task performance. We discuss how humans move between modes of AI use by increasing, reducing, or reinforcing their uncertainty tolerance. We conclude by discussing implications for the design of AI systems, policy into delegation in joint task performance, as well as use of data to improve learning from uncertainty.

2. Learning from uncertainty to use AI systems

The phenomenon of interest in this study is the ways by which humans learn from uncertainty when jointly performing a task with AI systems. We are particularly interested in the joint performance of *high uncertainty tasks such as medical and criminal justice decision making* whereby the margin for error is close to zero due to the potentially devastating consequences for human lives (Hartmann & Wenzelburger, 2021; Jussupow et al., 2021; Lebovitz et al., 2022). We focus on how humans learn from uncertainty that may emerge because of limited information (e.g., identifying a bollard in the road ahead) or because of events that are difficult to predict (e.g., predicting other drivers' behavior in the road). Specifically, we focus on both uncertainty that can be mitigated with more information,

[2] Other research examining decision-making under uncertainty has also pointed at varied heuristic systems employed by humans, based on individual preferences (Kahneman et al., 1982; Kahneman et al., 2021).

and uncertainty that cannot be mitigated because it lies outside the control of individual agents (Griffin & Grote, 2020; Packard & Clark, 2020).

In the first type of uncertainty, an individual actor has no information about what happens (and is going to happen), but is able to calculate the probability of risk involved (i.e., this is referred to as a "known unknown"); in the second type of uncertainty an individual actor knows neither what happens nor can they calculate the statistical probability of risk involved (i.e., this is referred to as an "unknown unknown") (Hartmann & Wenzelburger, 2021; see also Mousavi & Gigerenzer, 2014). Whereas known unknowns can be mitigated, unknown unknowns cannot because they are completely random. In a dynamic environment with multiple actors involved such as driving a car in a city, one actor's reaction to uncertainty is tightly coupled to those of other actors such that even a small error (e.g., accelerate instead of reduce speed) can rapidly escalate into further errors with potentially devastating outcomes for all actors involved (Perrow, 1999). Thus, actors can face both mitigable and unmitigable uncertainty simultaneously (Packard & Clark, 2020). For example, one actor avoiding a bollard in the road (a known unknown) can enter the trajectory of another actor and generate an unknown unknown for the second actor and potentially a fatal accident for both.

Humans learn from such uncertainty by engaging in exploration and exploitation in their task performance (Gershman, 2018; Gottlieb & Oudeyer, 2018; March, 1991; Walker et al., 2019). This research examines whether humans exploit existing options to perform the task at hand or explore information about alternative options for possible long-term gains. Humans learn over time, "but choices must be made between gaining new information about alternatives and thus improving future returns (which suggests allocating part of the investment to searching among uncertain alternatives), and using the information currently available to improve present returns (which suggests concentrating the investment on the apparently best alternative)" (March, 1991: 72). Specifically, both big (e.g., how to select a route for a destination) and small decisions (e.g., how to overtake a car) can benefit from exploration of alternatives before committing to and exploiting the benefits of a particular choice. By going through this process, humans engage in learning, whereby they sample and search information relevant for a task, such as looking at relevant stimuli while driving, and conferring value to that information (Cohen et al., 2007). Value is dependent on feedback accumulated from predicting a future state in the task (e.g., take a right turn while driving) and from prediction errors that are critical for learning. While actors may have options available with high value, instead of exploiting those, they may explore other alternatives with the prediction belief that those may generate more value in the long-term. This learning process reflects similar heuristics found in theories of satisficing (Simon, 1955) and in prospect theory (Kahneman, Slovic, & Tversky, 1982).

Key to our analysis is that learning from uncertainty is very much dependent on the *uncertainty tolerance* of individual actors (Griffin & Grote, 2020). For example, actors who are intolerant of uncertainty are likely to avoid or seek to reduce uncertain situations; whereas actors who are tolerant of uncertainty tend to be more explorative of uncertain situations (Cohen et al., 2007). Uncertainty tolerance has been found to influence hazardous attitudes and the safety operation of cars by human drivers, for example (Borowsky, Shinar, & Oron-Gilad, 2010; Sagberg & Bjørnskau, 2006). While these studies have focused solely on human performance, we are interested in investigating the dynamic and interdependent relationship between how humans learn and how such learning impacts their joint task performance with AI systems. Unlike previous research on algorithm aversion (Burton et al., 2020; Castelo et al., 2019; Dietvorst et al., 2015; Filiz et al., 2023; Reich et al., 2022), we are interested to understand how actors learn to jointly perform a task with an AI system both in *in real-time* and *over time*. We place emphasis on how actors learn to adjust their uncertainty tolerance by delegating tasks to the AI system and reinforcing their exploration or exploitation choices (e.g., increasing exploitation of automation or exploring alternatives depending on returns from joint task performance).

3. Methods

In this paper, we conduct phenomenon-focused qualitative research (Gkeredakis & Constantinides, 2019; Monteiro, Constantinides, Scott, Shaikh, & Burton-Jones, 2022) on Tesla's autonomous driving systems (ADS) using secondary data. We focus our analysis on examining how humans learn from uncertainty and how such learning impacts their joint task performance with AI systems.

3.1. Data collection

Research examining the performance of ADS often utilize accident report data and crash investigations (Favarò, Nader, Eurich, Tripp, & Varadaraju, 2017; Liu, Wang, Wu, Glaser, & He, 2021) and simulation tests in controlled (Bauchwitz & Cummings, 2020; Du et al., 2019; Eriksson & Stanton, 2017) and real-time environments (Banks, Eriksson, O'Donoghue, & Stanton, 2018; Eriksson, Banks, & Stanton, 2017). Accident report data and crash investigations are useful in understanding why accidents take place, under what conditions and how frequent, while identifying the prevalence of human error (whether by operators of autonomous systems or other human drivers) versus errors attributed to autonomous systems. Simulation test data in real-time environments are useful in understanding the behavior of humans while interacting with autonomous systems, including delegating tasks to them and monitoring their performance. Warnings and alerts are issued at close-collision incidents monitored by the designers of the simulations which provide rich data into the joint human-AI task performance.

We use three sources of data, (a) crash investigation reports provided by the US National Highway Traffic Safety Administration

(NHTSA) and the US National Transportation Safety Board (NTSB) involving Tesla's autonomous driving systems (ADS), (b) published reports on formal simulation tests of Tesla's ADS and (c) informal simulation tests by amateur drivers activating Tesla's ADS in major cities and recorded on YouTube. We also registered in the Tesla Owners Online forum[3] to search for relevant discussions on the use of Tesla's ADS. Here, we were able to identify such discussions as "Road to Full Self-Driving (FSD)", "Autopilot OFFICIAL THREAD" and "Autopilot vs Enhanced Autopilot", among others, that provided Tesla owners with insights into their experiences with different types of ADS. From these discussions, we were able to identify reactions to crash investigations, thus, connecting to our first source of data. We were also able to identify Tesla owner enthusiasts that activated ADS in their cars and recorded their experiences on YouTube, thus, connecting to our second source of data.

The selected data are not meant to be exhaustive, but rather illustrative of the range of interactions between humans and AI systems in jointly performing the driving task. The data reflect a longitudinal field experiment with Tesla ADS by a collective of humans, including Tesla owners, but also researchers, traffic safety and control officers, accident investigation teams, transportation boards and other related stakeholders. We examine these experiments to derive insights into how humans learn from uncertainty both in real-time and over time, focusing specifically on how Tesla drivers learn. Such learning feeds into Tesla's cloud servers that make it into updated versions of ADS. Thus, human learning impacts joint human-AI task performance.

We identified both successful and unsuccessful (e.g., running into accidents, humans intervening and taking over control of the car) learning from uncertainty across a range of driving scenarios. Specifically, the crash investigation reports focus on highway driving, with humans delegating control over to Tesla's ADS; the formal simulation reports focus on driving in controlled environments or on the highway, but always with a safety driver in the passenger seat; finally, the YouTube videos focus on major city driving by amateur drivers. We focus solely on data that support a comparable analysis of how humans learn from uncertainty in their joint task performance with Tesla's ADS. A summary description of our data is provided in Table 1.

3.2. Data analysis

We segmented all data into units of text relating to human-AI joint performance of the driving task. Following previous research in autonomous systems, we initially coded data based on the actions and reactions of human drivers to ADS (Banks, Stanton, & Harvey, 2014; Endsley & Kaber, 1999; Parasuraman, Sheridan, & Wickens, 2000). For example, we focused on the actions of humans such as *activating* Tesla's ADS, *monitoring* the performance of the ADS, *evaluating* such performance, and *responding* (if needed) with interventions such as, taking over control of the car. We also focused on the actions of Tesla's ADS such as *monitoring* whether humans were attentive to the driving task and generating visual and auditory alerts and *evaluating* the responsiveness of humans to these alerts (Du et al., 2019).

Next, we refined our analysis by using a hybrid of top-down (theory-driven) and bottom-up (data-driven) approach to understand how humans learned in real time and over time, while exhibiting different levels of uncertainty tolerance. In particular, we used our knowledge from the literature on algorithm aversion (Dietvorst et al., 2015; Filiz et al., 2023; Reich et al., 2022) and the uncertainty tolerance of individual actors (Griffin & Grote, 2020) to theorize how human drivers jointly performed the driving task with Tesla's ADS, while leveraging insights from the secondary data collected. This coding and analysis enabled us to inductively derive four aggregate theoretical themes as summarized in Table 2 below. These themes correspond to two temporal windows, namely, real-time learning from uncertainty with high and low uncertainty tolerance, and learning from uncertainty over time again while exhibiting varied tolerance to uncertainty. Data on real-time learning focused on actual use of Tesla's ADS, whereas data on learning over time focused on reflections of human drivers on their use of Tesla's ADS over different versions, comparing improvements and shortcomings between software updates.

In addition to the four aggregate themes, we also placed emphasis on the transfer of learning between the two temporal windows. In particular, we wanted to understand when human drivers increased, reduced or reinforced their uncertainty tolerance after using Tesla's ADS and learning from their own performance and the performance of the AI system. We discuss this meta-level analysis of learning transfer in the discussion section when we present our theoretical model.

4. Analysis

As background for our analysis, we first provide a description of Tesla's ADS system. Next, we proceed to present our analysis of the four themes of how humans learned from uncertainty.

4.1. Tesla's autonomous driving systems (ADS)

On August 20th 2021, Tesla hosted "Tesla AI Day"[4] to introduce their ADS technology stack and Autopilot in detail. This enabled us to better understand how Autopilot performs the driving task, which was essential for our analysis. Andrej Karpathy, the then senior director of Tesla AI, explained that Tesla is "*effectively building a synthetic animal from the ground up. So, the car can be thought of as an animal; it moves around it senses the environment and, you know, acts autonomously and intelligently*." He elaborated that, "*the brain of the Autopilot*" depends on a "*synthetic visual cortex*" that resembles the biological visual cortex of an animal. This is why computer vision is

[3] See https://www.teslaownersonline.com/

[4] See relevant video here https://www.youtube.com/watch?v=j0z4FweCy4M

Table 1
Data involving Tesla's ADS.

Type of data	Short description
Crash Investigation Report (CIR): CIR 1	Fatal accident of a 2015 Tesla Model S 70D sedan with the right plane of a 2003 utility semi-trailer. Source: https://crashstats.nhtsa.dot.gov/Api/Public/Publication/812481
Crash Investigation Report (CIR): CIR 2	Accident of a 2014 Tesla Model S P85 with a parked fire truck. Source: https://www.ntsb.gov/investigations/AccidentReports/Reports/HAB1907.pdf
Crash Investigation Report (CIR): CIR 3	Accident of a 2017 Tesla Model X P100D with a non-operational crash attenuator, with subsequent collisions with two other vehicles, a 2010 Mazda 3 and a 2017 Audi A4. Source: https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR2001.pdf
Crash Investigation Report (CIR): CIR 4	Fatal accident of a 2018 Tesla Model 3 with a 2019 truck-tractor in combination with a semitrailer. Source: https://www.ntsb.gov/investigations/AccidentReports/Reports/HAB2001.pdf
Simulation Report (SR): SR 1	Controlled driving of a 2018 Tesla Model S P90 in public roads in the UK involving twelve participants between the ages of 20 and 49 years in the presence of a safety driver. Source: Banks et al. (2018)
Simulation Report (SR): SR 2	Controlled driving of three 2018 Tesla Model 3S in a public highway or at the North Carolina Center for Automotive Research (NCCAR), a closed test track facility, in the presence of safety driver. Source: Bauchwitz and Cummings (2020)
YouTube Video (YTV): YTV1	Use of Tesla's Full Self Driving (FSD) Beta 10.12 software version in San Francisco, California. Source: https://www.youtube.com/watch?v=fwduh2kRj3M
YouTube Video (YTV): YTV2	Use of Tesla's Full Self Driving (FSD) Beta 10.10 software version in Detroit, Michigan. Source: https://www.youtube.com/watch?v=BeoFYPn_Beo
YouTube Video (YTV): YTV3	Use of Tesla's Full Self Driving (FSD) Beta 10.10 software version in San Jose, California. Source: https://www.youtube.com/watch?v=sbSDsbDQjSU
YouTube Video (YTV): YTV4	Use of Tesla's Full Self Driving (FSD) Beta 10.12 software version in Southampton, Pennsylvania. Source: https://www.youtube.com/watch?v=z3wM3xdOdI0
YouTube Video (YTV): YTV5	Use of Tesla's Full Self Driving (FSD) Beta 10.10 software version in Rochester, Michigan. Source: https://www.youtube.com/watch?v=Kf3t9gNOl98
YouTube Video (YTV): YTV6	Use of Tesla's Full Self Driving (FSD) Beta 10.12 software version in San Francisco, California. Source: https://www.youtube.com/watch?v=Vz-SyKC9qnM
YouTube Video (YTV): YTV7	Use of Tesla's Full Self Driving (FSD) Beta 10.12 software version in San Francisco, Brooklyn and Middle Tennessee. Source: https://www.youtube.com/watch?v=D_SPymCay18
YouTube Video (YTV): YTV8	Use of Tesla's Full Self Driving (FSD) Beta 10.12 software version in Salt Lake City Source: https://www.youtube.com/watch?v=ZLBR39RcyiU

key in the Tesla ADS, since it captures, just like the biological visual cortex of an animal, raw images with the help of eight cameras that are positioned around the vehicle. These eight cameras generate a three-dimensional representation of objects captured with three-dimensional positions of lines, edges, curbs, traffic signs, traffic lights, cars and their orientations, depth, and velocities. Unlike other ADS, such as Google's Waymo, Tesla does not use any other sensors such as LIDAR and ultra-sonic sensors. Tesla's technological shift from LIDAR to computer vision was an effort to differentiate itself from its competitors but also to address errors in earlier models that led to car crashes (see CIR 1–4 in our analysis). Even customers who owned earlier models (manufactured prior to 2021) had their LIDAR sensor disabled through software updates.

The input to Tesla's cameras is raw video data (1280 × 960 12-Bit @ 36 Hz). These data are processed by RegNet (regular network structures) designed with residual neural network blocks (ResNet).[5] This neural network generates outputs of several features at different resolutions and at different scales, starting with very high-resolution features with a very low channel count and moving all the way to low resolution but high channel counts. So, the neural network has neurons that scrutinize the detail of raw video data and capture these data in its broader context. All features are next processed by a bi-directional feature pyramid network (BiFPN)[6] that uses multiple scales to weigh the relevance of each feature. The BiFPN enables Tesla's ADS to engage in multiple tasks, not just object detection, but also traffic light recognition and lane prediction. Hence, with the RegNet neural network and the BiFPN acting as a "backbone" infrastructure that branches into a number of "heads" responsible for multiple, interdependent tasks, the key objective of Tesla's ADS technology stack is essentially to integrate a lot of different technologies which they can then decouple, upgrade and validate for each task. At the same time, feature training is end-to-end, meaning that the ADS learns holistically to perform all tasks.

This architecture allows Tesla's ADS to plan a path *jointly* with the explicit plans and interventions of the human driver, as well as the plans and trajectories of other cars in the road, as Ashok Elluswamy, director of the Autopilot software, explained at the Tesla AI Day presentation. This creates a rather complex problem that requires a hybrid approach to searching for solutions and continuously optimizing them. Accordingly, Autopilot uses a combination of reinforcement learning algorithms together with a value function (e.g., minimizing 'cost' such as travel time and distance), which aims at maximizing three rewards or objectives: the safety, comfort, and efficiency of the car.

Based on the Tesla website,[7] Autopilot is only available on cars built on September 2014 or later. Cars without this software, but which are equipped with the necessary hardware, can be fitted with Autopilot, Enhanced Autopilot or Full Self-Driving Capability

[5] See https://arxiv.org/abs/2003.13678 for a detailed explanation of this type of neural network.
[6] See https://arxiv.org/abs/1911.09070 for a detailed explanation of how these feature pyramid networks work.
[7] See https://www.tesla.com/support/autopilot

Table 2
Data analysis and themes.

Sample quotes	Coding	Aggregate Theoretical Themes	
"The car is obviously having trouble with that. … it's just gonna try to turn left into that person's driveway, so unfortunately, I'm going to report this one so the team sees that… Sometimes there's garbage trucks you know. People would know about that. Nobody's gonna get confused by this. But I guess in terms of the car software it's kind of an edge case it hasn't seen that enough to know like okay there's no room I just need to wait." [YTV 5, 11:48–12:23 min]	AI system does not react early enough to hazards or performs inappropriately because of "edge cases" that would be normal for humans, causing an intervention by human driver	*1. Intervention into AI system performance to mitigate uncertainty* Humans with *low* uncertainty tolerance explore automation by AI systems in real-time, but are continuously monitoring and intervening in AI task performance while exploiting their own knowledge to mitigate uncertainty.	Learning in real time
"It's still kind of my biggest complaint right now, is that, you know, phantom breaking … since we switched to vision it's been a problem for a while but it's really not getting much better … so when you're going above maybe 50 or so on a road like this it's just you get a lot of these little breaks here and there and it's super annoying… the car will go from 60 to 55 or 60 to you know 50 and you have enough time to touch the accelerator and all this …". [YTV 5, 15:12–15:57 min]	AI system performs the task inscrutably (e.g., phantom breaking[a]) even though there is no real hazard, causing an intervention by human driver		
"Wow! That was too aggressive and like I had to take over. I had to take over because it was going too far to the left." [YTV 2, 8:29–8:33 min] "*Oh God! Going towards those bollards again. I really don't like that*" [YTV 3, 7:47–7:49 min]	AI system incorrectly recognizes objects (e.g., lanes, bollards) causing an intervention by human driver		
"During the 5 seconds leading up to the crash, the Tesla SUV accelerated from about 63 mph to nearly 71 mph at impact. During the approach to the crash attenuator, the FCW system did not provide an alert and the AEB did not activate" [CIR 3, p. 16]	AI system does not recognize objects in the environment and human driver is completely disengaged. Uncertainty is not mitigated leading to an accident	*2. Overconfidence and automation complacency in AI system performance (uncertainty not monitored)* Humans with *high* uncertainty tolerance explore automation by AI systems in real-time, but become disengaged from the task themselves. This leads to overconfidence in AI capabilities and automation complacency. Uncertainty cannot be mitigated leading to accidents.	Learning in real time
"*After 60s [seconds] [of hands free driving], a visual warning was presented in the HMI [human machine interface] that instructed them to place their hands back on the wheel. This went completely unnoticed for 15 s before an auditory tone was sounded. At this point, the driver appeared confused … The ability of the driver to respond in this situation was weakened as their awareness surrounding system state was compromised.*" [SR 1, p. 140]	AI system responds to human driver inactivity by generating alerts. Human driver appears confused, system is compromised. In the absence of a safety driver, uncertainty would have become unmitigated leading to an accident.		
"The Autopilot system and collision avoidance systems did not identify the crossing truck as a hazard and did not attempt to slow the car. In addition, the driver did not receive an FCW alert, and the AEB system did not activate." [CIR 4, p. 14]	AI system does not recognize hazard and does not alert human driver on time. Uncertainty is not mitigated leading to an accident		
"The Tesla's multiple ADAS and CA technologies, including Autopilot and FCW, were functional at the time of the crash. … their limitations made it difficult for the system to recognize this specific crash due to the cross-path configuration of the involved vehicles' trajectories and the overall physical characteristics of the intersection/roadway." [CIR 1, p. 25 and p. 2]	AI system fails to recognize hazard and does not alert human driver on time. Uncertainty is not mitigated leading to an accident		

(*continued on next page*)

Table 2 (*continued*)

Sample quotes	Coding	Aggregate	
		Theoretical Themes	
"I think FSD beta is currently achieving about 90% accuracy in decision making. Maybe is 95%. Maybe it is 80%. But I'll give it the benefit of doubt, and let's say it is 95% accuracy. That means 5 times out of every 100 it is making an incorrect decision. Since it is making hundreds of decisions [per]second, errors are far too frequent. I don't think I've ever completed a FSD Beta drive without disconnecting, usually multiple times. … And I don't think I've seen anything more than a 1–2% improvement in performance over the last 6 months and 8 versions of FSD Beta I'm on." TetonTesla, (Tesla Owners Online Forum)[b]	Human driver is frustrated with the lack of progress in full driving automation and the errors made by the AI system. Human driver has to always take over the task from the AI system.	*3. Fear & control over AI system performance (perceived inability of the AI System to mitigate uncertainty by itself)* Humans with *low* uncertainty tolerance are limited in their exploration of automation by AI systems over time. They fear the probability of uncertainty and seek to control for errors in task performance by themselves. This leads to little to no learning in joint task performance	Learning over time
"*My FSD beta is now safely stored where it belongs; in the novelties section between autopark and summon. For the first time in 30 days, I went for a drive this morning and drove just like I wanted to. What a joy!! Good riddance FSD beta.*" FRC, (Tesla Owners Online Forum)[c]	Human driver questions the safety of the AI system reverting instead to manual driving.		
"I think the concern is with the latter part 'while I monitor'. People are finding ingenious ways around the built-in monitoring. As software gets more advanced the fear is that it shouldn't be out there unless there's a 100% way to ensure driver attention (and there isn't) or the software is actually ready (it won't be)." Shareef777(Tesla Owners Online Forum)[6]	Human driver questions the ability of both humans and the AI system to monitor one another. There is fear of errors in AI system performance and the ability to mitigate uncertainty.		
"…version 10.10… noticed some big improvements … I will start off with some of the good stuff first … It seems to be able to predict the actions of vulnerable road users like pedestrians and bicycles with much greater confidence and earlier predictions than before…" [YTV 6, 7:26–8:00 min]	AI system learns to predict actions with "much greater confidence than before", thus, can mitigate uncertainty by itself.	*4. Advance AI system performance (AI System mitigates uncertainty by itself)* Humans with *high* uncertainty tolerance explore automation by AI systems over time, while exploiting their own knowledge of the task. This leads to joint learning in task performance, advancing the ability of the AI system to mitigate uncertainty by itself.	Learning over time
"*It seems to be predicting other vehicles and tensions as well and oftentimes seems like it's planning ahead for its manoeuvrers. We have a right turn to make ahead, and you can see it highlighting some of the vehicles to our right in blue on the visualization and predicting their past to try to find a gap to fit into while we make the lane change.*" [YTV 6, 8:50–9:34 min]	AI system learns to plan for its maneuvers based on predictions of other vehicles' positions, thus, can mitigate uncertainty by itself.		
"*What an awesome update 10.12 ended up being. It includes new training data from over 250,000 real world video clips and an absolutely massive change log …It learns from actual experiences and improves itself just like we do*" [YTV 1, 11:27–11:36 min] "*These streets are not easy and when the Autopilot is making mistakes it's finding ways to correct them without making awkward situations most of the time."* [YTV 1, 12:38–12:44 min]	AI system learns from "actual experiences" and corrects mistakes, thus, can mitigate uncertainty by itself.		

[a] "Phantom breaking" is sudden, unexpected, automatic breaking performed by Tesla's ADS. See https://www.washingtonpost.com/technology/2022/02/02/tesla-phantom-braking/

[b] Discussion Topic: FSD v10.11.2. https://www.teslaownersonline.com/threads/fsd-v10-11-2-2022-4-5-21.21498/#post-336965

[c] Discussion Topic: FSDBeta Megathread For all FSD Beta Discussions. https://www.teslaownersonline.com/threads/fsdbeta-megathread-for-all-fsd-beta-discussions.18878/page-89#post-322871

(FSD) through the Tesla app. Enhanced Autopilot and FSD are additional software modules that can be purchased and added to the standard Autopilot module. Together, they incorporate such functionalities as traffic-aware cruise control, autosteer, navigation and auto lane change, forward and side lane collision warning, automatic emergency braking and lane departure avoidance. In our analysis, we focus on early versions of Autopilot fitted in 2015–2018 models and FSD beta versions fitted in 2021 models. Collectively we refer to these as Tesla's Autonomous Driving Systems (ADS).

4.2. Learning from uncertainty in driving

According to the National Highway Traffic Safety Administration (NHTSA),[8] as of June 2022, there are approximately 830,000 Tesla cars in the US with ADS installed. Tesla claims that over 9 billion miles[9] have been driven with ADS activated. As described above, Tesla uses the data from this fleet of cars to understand how its ADS performs and to train the system to avoid accidents. One Tesla owner, known as AI DRIVR and regularly posting his driving tests on YouTube to reach his 93 k subscribers said [YTV 1, 11:12–11:41]:

> Honestly, I never realized how many nuances there are to driving until I started using the [FSD] Beta. It's going to be very difficult to create a generalized driving machine, obviously much more difficult than Elon initially thought. But I think data collection from other cars making mistakes in the real world is the only way to make it a reality.

As this quote shows – and further discussions in the Tesla owners' forum confirm – even Tesla enthusiasts were reflective of their use of Tesla's ADS, understanding that joint learning is ongoing. In examining how amateur Tesla drivers activated ADS in their cars and jointly performed the driving task we arrived at four aggregate themes. Each of these themes point at the ways by which humans learn to perform the driving task with ADS activated both in *real time* and *over time.* These themes also highlight instances of varied levels of delegation and control between human drivers and ADS, due to the human drivers' *low* or *high* uncertainty tolerance. Our presentation of the findings below follows the structure and content of Table 2.

4.2.1. Intervention into AI system performance to mitigate uncertainty

Our analysis pointed at one theme, whereby human drivers with *low* uncertainty tolerance explore automation by AI systems in real-time, but are continuously monitoring and intervening in AI task performance while exploiting their own knowledge to mitigate uncertainty. One Tesla Owner, Kevin Smith, with more than 5000 miles of FSD test drives[10] said [YT7 1:49–2:01]:

> With it being a beta, you know, I've seen that any time the car could just make a mistake, and I have to be ready for that. Right now, my stress levels go up, not down from using full self-driving, but that's the cost of making it better.

Indeed, despite their automation enthusiasm, these amateur Tesla drivers had no misconception of the dangers of autonomous driving and were always alert and attentive to the driving task, ready to intervene if needed. One Tesla owner, known as Black Tesla, noted in one of his test drives [YTV 4, 4:33–4:50 min]:

> Don't be afraid to disengage … be overly cautious like a child, like a teen driver. Just don't be afraid to disengage and get yourself into a weird situation. If you feel like the car is doing something or is not doing something the right way jump in, take over immediately.

At the same time, these Tesla drivers were putting the ADS through extensive tests, commenting with great detail on their task performance, while continuing to be vigilant of the driving task, as the following example shows [YTV 1, 8:14–8:47 min]:

> After this left, we have to make an immediate right and there's a truck we have to account for. I would have just slowed down early and let the truck go first but Autopilot does not do this. Seems like it wants to pass it or something and get side by side with it and then has to make awkward last minute lane changes to get over. It's also treating this yielding sign like a stop sign. There was nobody behind us, so it was fine, but this bit definitely didn't feel as natural as it had been before. But still no disengagements or interventions so far, so doing a pretty great job overall.

As this quote shows, the performance of the ADS was not perfect. The human driver, however, supervised the task performance of the ADS, while being situationally aware and ready to intervene if needed as the latter performed autonomously. Indeed, in some cases, the human driver had to intervene because the ADS was not reacting early enough or performing inappropriately. This Tesla owner, known as Dirty Tesla, explains what happened in one test drive [YTV 5, 11:48–12:23 min]:

> The car is obviously having trouble with that … it's just gonna try to turn left into that person's driveway, so unfortunately, I'm going to report this one so the team sees that … Sometimes there's garbage trucks you know. People would know about that. Nobody's gonna get confused by this. But I guess in terms of the car software it's kind of an edge case it hasn't seen that enough to know like, okay, there's no room. I just need to wait.

As this human driver explains, a garbage truck stopping in the middle of the street for a short period of time before moving forward seems to be an "edge case" for the ADS. Given the lack of traffic signals, the ADS had difficulty in identifying the difference between a person's driveway and the continuation of the street. The ADS sought to maximize efficiency – in this case minimize the time of the trip. However, this led to a compromise in the safety of the car by attempting to go into a person's driveway in order to overtake the garbage truck. So, the human driver was confused as to why the ADS was attempting a risky path and with what implications for his own safety and potentially of others that happened to be in the driveway. In the end, the human driver was quick to intervene to stop the ADS from changing its path. As this example shows, the ADS created uncertainty for the human driver and potentially for other drivers in the

[8] See https://static.nhtsa.gov/odi/inv/2022/INOA-EA22002-3184.PDF
[9] See https://www.tesla.com/VehicleSafetyReport
[10] Kevin Smith's channel https://www.youtube.com/@spleck615

road. The human driver intervened to avert a small error escalate into a devasting outcome such as an accident. This was an 'edge case' causing the ADS to make an error that it could not autocorrect, but which was relatively easily corrected by an alert human driver, who then reported the case back to Tesla.

Some of the errors made by ADS were due to the underlying technology stack, especially its reliance on computer vision to detect objects and then engage in a number of other tasks, such as acceleration, deceleration and path planning. One of the highest reported problems with the performance of Tesla's ADS is what is referred to as "phantom breaking". Phantom breaking happens when Tesla's computer vision technology perceives shadows or optical illusions in the road as real objects because it lacks depth that would be required to validate the authenticity of the object. This problem was so extensive that the NHTSA opened up an investigation of 758 reports of unexpected brake activation.[11] The Tesla owner known as Dirty Tesla explained in one of his driving tests [YTV 5, 15:12–15:57 min]:

> It's still kind of my biggest complaint right now, is that, you know, phantom breaking … since we switched to vision it's been a problem for a while but it's really not getting much better … so when you're going above maybe 50 or so on a road like this it's just you get a lot of these little breaks here and there and it's super annoying … the car will go from 60 to 55 or 60 to you know 50 and you have enough time to touch the accelerator and all this but … it's been a problem for so long and it seems like a vision thing where the car just can't see far enough for the speed it's going.

Phantom breaking refers to ADS performance that is inscrutable to humans. Such inscrutable performance happens unexpectedly and causes human drivers to intervene (e.g., touch the accelerator) to correct errors in driving. In addition, we also observed cases where human drivers had to completely take over control of the car because of imminent danger. For example, one Tesla owner, known as AI Addict, while driving in downtown San Jose experienced quite erratic behavior by the ADS, changing lanes with no apparent reason, making very tight turns and accelerating into bollards:

> Oh, this is, no, it found it … but it almost took us to the wrong lane [left turn into the wrong lane].
>
> [YTV 3, 2:22–2:29 min]

Other Tesla owners also experienced such erratic behavior by the ADS, as it went too far into the opposite lane and changed lanes without signalling:

> Wow! That was too aggressive and like I had to take over. I had to take over because it was going too far to the left.
>
> [YTV 2, 8:29–8:33 min]

> Changing lanes erratically for no reason here. I have a hand on the yoke to take over as needed.
>
> [YTV 4, 8:09–8:15 min]

Also, in the comments section of YTV 2, other Tesla owners elaborated:

> It would be interesting to repeat a failed route to see if the performance is consistent. I have found that bad lane selection is very consistent, and I think it's due to errors in the maps used for path planning. On my test routes, the car makes the same irrational lane changes, often displaying a message "Changing lanes to follow route" when no change of lane is necessary.
>
> (YouTube profile: Clara Smith)

> We've noticed across a few different cars that use Google Maps, not only Tesla, that turn by turn directions have been doing strange routing all of a sudden the last week or so. I think there is some Google Maps bugs at the moment which may not be helping with FSD.
>
> (YouTube profile: PyroJoe)

Evidently, with every new update, the ADS had to run through different routes to learn to correct errors such as unidentified lanes, but also errors in maps. While these routes would appear normal to human drivers, these appeared as edge cases to the ADS. The low uncertainty tolerance and full attention of these human drivers in exploring ADS performance was paramount to reacting quickly, correcting errors and mitigating uncertainty. Through such experimentation and human intervention across a number of variations of collected data (e.g., lane changes on marked vs unmarked lanes), Tesla's RegNet was able to correct errors, by saving all instances in real-time to later upload and add to new versions of the software.[12] Tesla uses such data not only for object detection, but also for path planning, and for predicting different scenarios involving other drivers, cyclists, or pedestrians. We discuss such learning in Section 4.2.4.

4.2.2. Overconfidence and automation complacency in AI system performance (uncertainty not monitored)

Our analysis pointed at a second theme, whereby humans with *high* uncertainty tolerance explore automation by AI systems in real-time, but become disengaged from the task themselves. This leads to overconfidence in AI capabilities and automation complacency. In

[11] See https://www.theverge.com/2022/6/3/23153241/tesla-phantom-braking-nhtsa-complaints-investigation

[12] See https://www.youtube.com/watch?v=Ucp0TTmvqOE at point 1:59:38 when Andrej Karpathy explains how RegNet learns from the fleet of Tesla cars driven by amateur drivers.

the previous subsection, human drivers were always alert, attentive to the driving task and ready to intervene and even to take over full control of the car when ADS seemed to perform erroneously and run the risk of an accident. In the following crash investigation and controlled simulation reports human drivers were completely inattentive to the driving task. As a consequence, uncertainty could not be mitigated leading to accidents.

In the controlled simulation reports, once ADS was activated to take control of the Tesla cars, human drivers exhibited what previous research has called "automation complacency," by losing situation awareness of the task (Endsley, 2017) and becoming too complacent in detecting failures in automation (Parasuraman & Manzey, 2010). For example, in SR 1 the following behavior was recorded by the cameras mounted to record the human driver:

> After 60s [of hands free driving], a visual warning was presented in the HMI [human machine interface] that instructed them to place their hands back on the wheel. This went completely unnoticed for 15s before an auditory tone was sounded. At this point, the driver appeared confused … The ability of the driver to respond in this situation was weakened as their awareness surrounding system state was compromised.
>
> [SR 1, p. 140]

Because of their loss of situated awareness, the human driver appeared confused when the ADS generated both visual and auditory alerts for the former to place their hands back on the steering wheel. The human driver simply had no idea why the alerts were generated and what they were meant to do to respond. This problem has been well reported in other car simulation experiments that examined other types of automation (Nilsson, Strand, Falcone, & Vinter, 2013; Samuel, Borowsky, Zilberstein, & Fisher, 2016). These experiments showed that, precisely because they lose situation awareness, human drivers experience uncertainty regarding the potential hazard, including how to respond once they take over from the ADS.

The crash investigation reports provide plenty of evidence of this – in some cases, fatal – relationship between loss of situation awareness and the learning from uncertainty. For example, in the second crash investigation report, where a Tesla car run into a stationary fire truck, the human driver was found to be completely inattentive to the driving task, something that left them with no knowledge of the pending danger. By extension, they also could not assess the probability of an accident, nor could they evaluate the best course of action to avoid it:

> In this crash, the driver's lack of braking and steering in response to the stopped fire truck, his statement that he never saw the fire truck, and his potential in-vehicle distractions (bagel, cup of coffee, radio) all suggest that the driver was not attending to the driving task before the crash.
>
> [CIR 2, p. 13]

Because of the human driver's inattention to the driving task, all the challenges faced by the ADS in the Tesla cars discussed in the previous sections became aggravated. Under no supervision and faced with uncertainty (i.e., stopping or changing lanes), the ADS were found to be unable to correct the errors to avoid accidents. For example, in the second crash investigation, the ADS was faced with the problem of not detecting the crash attenuator and accelerating into it:

> During the 5 seconds leading up to the crash, the Tesla SUV accelerated from about 63 mph to nearly 71 mph at impact. During the approach to the crash attenuator, the FCW [forward collision warning] system did not provide an alert and the AEB [automatic emergency braking] did not activate.
>
> [CIR 3, p. 16]

The crash investigation puts the probable cause of the accident down to "*the Tesla vehicle's ineffective monitoring of driver engagement, which facilitated the driver's complacency and inattentiveness*" [CIR 3, p. ix]. What is also happening here is that Tesla's LIDAR technology did not recognize the attenuator as a stationary object in the Tesla's trajectory. The crash attenuator was damaged from a prior accident and the LIDAR technology was arguably not fit to identify it. Such cases were used by Tesla to eventually phase out LIDAR in 2021 and install computer vision only in later cars. A Tesla owner posted a video of a similar incident, where Tesla's LIDAR technology failed to recognize the concrete barrier on the highway running straight into it.[13] Fortunately, in that case, the human driver was alert and attentive to the driving task enabling him to quickly take over and avoid an accident. In the comments section of the video, another Tesla driver commented:

> Angle was too shallow for the radar to detect, and the divisions between the blocks confused the cameras and ultrasonic proximity sensors. That's my guess at least. They talked about this kind of thing in their Autonomy Day presentation, the more the cars are exposed to weird edge cases like this, the better equipped they'll become to deal with them.

Despite the problems with Tesla's ADS technology stack (see also Sections 4.2.3 and 4.2.4 for a discussion of how humans learned over time while using newer versions of the ADS), the key difference between the human drivers in the previous section and the human drivers in these crash investigation reports was that the latter were found to exhibit high uncertainty tolerance that made them complacent to automation. Automation complacency translates to overconfidence in ADS performance and complete inattention to the task at hand.

As evident from the crash investigation reports, these were incidents whose trajectories were very difficult to predict and yet there

[13] The video is posted here https://www.youtube.com/watch?v=fKyUqZDYwrU.

was exploration of risky options and no exploitation of confident knowledge about the task. In particular, in the two fatal accidents, the human drivers became so overconfident of the ADS performance that they fell asleep at the wheel (in CIR 1, the seat was adjusted "*to its rearmost track position, with the seatback slightly reclined*", p. 27) or were completely disengaged with the driving task, possibly occupied with other tasks (in CIR 4, the "*car driver, traveling at a recorded speed of 69 mph, did not apply the brakes or take any other evasive action to avoid the truck, which was crossing in front of him at about 11 mph*", p. 1). This was also true in the second simulation report whereby the three cars used exhibited "unsafe Autopilot behavior" that "would have likely led to an adverse event given a distracted driver" [SR 2, p. 25].

In both CIR 1 and CIR 4, the ADS in the two Tesla cars failed to recognize hazards in the road and did not alert human drivers on time to avoid the fatal accidents. In both fatal accidents, ADS were faced with extreme 'edge cases' involving complex physical characteristics. As reported in CIR 1:

> The Tesla's multiple ADAS and CA technologies, including Autopilot and FCW, were functional at the time of the crash … their limitations made it difficult for the system to recognize this specific crash due to the cross-path configuration of the involved vehicles' trajectories and the overall physical characteristics of the intersection/roadway.
>
> [CIR 1, p. 25 and p. 2]

Similarly, in CIR 4 it was reported that:

> The Autopilot system and collision avoidance systems did not identify the crossing truck as a hazard and did not attempt to slow the car. In addition, the driver did not receive an FCW alert, and the AEB system did not activate.
>
> [CIR 4, p. 14]

Remarkably, CIR 1 and CIR 4 are almost identical crashes that happened over the span of two years between 2016 and 2018. At that time, there was limited amateur driving and experimentation with Tesla's ADS feeding into Tesla's RegNet and, indeed, RegNet was not available. Thus, human learning from uncertainty and recording such learning into future versions of the ADS was impossible. This was also a time when ADS depended on LIDAR that according to Tesla was responsible for many errors, as discussed earlier. At the same time, the Autopilot version used in those early Tesla cars was never meant to be driven on highways and certainly not without human supervision. Our analysis shows that this absence of a human-in-the-loop was detrimental to monitoring uncertainty and correcting errors in joint task performance.

4.2.3. Fear and control over AI system performance (perceived inability of the AI system to mitigate uncertainty by itself)

Interestingly, these early fatal accidents had a negative impact on the uncertainty tolerance of some Tesla drivers. Specifically, our analysis points at a third theme of humans with *low* uncertainty tolerance being limited in their exploration of automation by AI systems over time. These humans fear the probability of uncertainty and seek to control for errors in task performance by themselves. This leads to little to no learning in joint task performance. Consequently, many Tesla owners never or rarely activate ADS, choosing to manually drive their car instead.

One YouTuber, named Snazzy Labs who mostly does tech reviews instead of Tesla test drives, posted one experiment with FSD highlighting its problems:

> I bought my model 3 in 2018 and I paid 8000 for the Autopilot suite as well as Tesla Full Self Driving but 2018 came and went and 2019 and 2020 and … Full Self Driving did not seem any closer…
>
> [YTV 8, 0:15–0:31 sec]

> Man, it really doesn't like being told what to do either. If you like push the gas and you're like "no go", it's like "oh, what are you doing?" Okay. It needs to turn right here but it's trying to get in this lane. I do not know why.
>
> [YTV 8, 11:38–11:50 min]

Interestingly, he tries to 'tell the car what to do', a paradoxical objective that goes against any effort towards automation. Unlike the test drives by other Tesla owners discussed in Section 4.2.1, that are systematic (run multiple times in the same and in different areas), the test drive performed by Snazzy Labs was a one-off and evidently with very low uncertainty tolerance to the point that ADS was not delegated to autonomously perform the driving task. Similarly, in the comments section of YTV 4, while commenting on the lane issue discussed in Section 4.2.1, a Tesla driver said:

> I'm still getting used to Autopilot and have only tried FSD beta a few times. I'm good until I encounter another car and then feel the need to take control.
>
> (YouTube profile: Roseanne Grindle)

Just like Snazzy Labs, the uncertainty tolerance of Roseanne Grindle is low. In another example, CNBC run an experiment[14] of three test drives, one in Brooklyn, another in San Francisco and another in Middle Tennessee. Out of the three, only the one in Middle Tennessee was driven by an experienced Tesla owner with more than 5000 miles of FSD test drives and who seemed to have no issues with FSD. The other two were evidently first-time testers of FSD and exhibited very low uncertainty tolerance. The experience was heavily criticized by many Tesla owners in the comments section of the video. For example:

[14] See https://www.youtube.com/watch?v=D_SPymCay18

> I use FSD all of the time on two lane roads and really like it. I know that it doesn't do well in certain situations, and I tend to be on high alert or I simply disengage at those times. This guy doesn't seem capable of making that leap. He just put it wherever he wants to and then says oh this thing doesn't work! He is unbelievable.
>
> (YouTube profile: Senor Dockman)

> Christ I actually went into this with an open mind. I've used FSD and this is not my experience at all.
>
> (YouTube profile: Lanre Ogun)

> If you want some real FSD content, go watch any of the [FSD] Beta tester YouTubers who have been doing it from the start, regular people who want to help the Beta learn and document its progress ... They show where its good, where its bad, and know how to handle it properly and not freak out whenever it does anything remotely unexpected.
>
> (YouTube profile: Kai Hyena)

The CNBC experiment was also criticized by other Tesla drivers on their own channels who were quick to note CNBC's biased representation of Tesla's FSD:

> What we see here in my opinion is pretty disingenuous. If you use the beta or have some experience with it you can probably figure this one out pretty quick… it's the human failing in this clip not the car.
>
> [YTV 6, 03:01–04:15 min]

The CNBC experiment makes a number of claims about the safety of Tesla FSD, many of which are true, but while building a case that automation is impossible. We also found evidence of this in the Tesla Owners Online Forum, whereby several Tesla drivers commented on how, despite several FSD Beta versions being released, there were still significant errors reported. Indeed, these Tesla drivers questioned the safety of FSD Beta, choosing instead to manually drive their Teslas, as one user with the pseudonym 'FRC' noted (Tesla Owners Online Forum)[15]:

> My FSD beta is now safely stored where it belongs; in the novelties section between autopark and summon. For the first time in 30 days, I went for a drive this morning and drove just like I wanted to. What a joy!! Good riddance FSD beta.

Interestingly, past studies have shown that when there are automation failures as in the examples of the fatal accidents discussed in the previous subsection, human drivers tend to re-evaluate their decision to delegate automation to AI systems (Castelo et al., 2019; Dietvorst et al., 2015; Reich et al., 2022). The key presumption of humans in this research is that AI systems make errors and that they cannot learn to correct those; in consequence, humans avoid using AI systems altogether. In other words, humans have very low levels of uncertainty tolerance with a direct impact on the long-term learning effects for both human and AI systems.

4.2.4. Advance AI system performance (AI system mitigates uncertainty by itself)

In contrast to the above, we found that Tesla drivers that actively engaged ADS learned from uncertainty, while also enabling the ADS to improve its performance. Specifically, a fourth theme in our analysis reveals that humans with *high* uncertainty tolerance explore automation by AI systems over time, while exploiting their own knowledge of the task. This leads to joint learning in task performance, advancing the ability of the AI system to mitigate uncertainty by itself. Going back to the CNBC experiment and the more experienced Tesla driver, he noted in relation to the lane issues we discussed in Section 4.2.1:

> So basically, before the 10.5 update and including 10.5, most of the issues that we're having were related to like turn lanes and odd street markings and things like that. With 10.12, all those little turn lane issues and things, those all kind of went away. So, they fixed those little things.
>
> [YT 7 11:41–11:58]

A clarification is needed here. As discussed earlier, through human experimentation and intervention, Tesla's RegNet was able to collect more data and correct errors, feeding new versions of the software to individual cars through cloud-based updates. However, such updates only include instances of learning bound to the physical experimentation of individual human drivers. In other words, human lessons learned in Middle Tennessee will only benefit other Tesla drivers and their cars in that location. Those lessons will not apply in Brooklyn New York, because the physical configuration of the roads and traffic, the nature of driving and other such features will impact how humans and ADS learn from uncertainty in that location very differently. So, even though the collective of human drivers and their Tesla cars benefit from cloud-based learning and updates, local learning from uncertainty will differ.

For those human drivers who consistently and systematically explored automation by ADS the learning effects in their joint task performance were significant. On some occasions, the ADS performed extremely well, being able to predict the behavior of other human drivers, as well as pedestrians and cyclists. In other words, the ADS learned to correct errors in their performance:

> It seems to be able to predict the actions of vulnerable road users like pedestrians and bicycles with much greater confidence than before.

[15] Discussion Topic: FSDBeta Megathread For all FSD Beta Discussions. https://www.teslaownersonline.com/threads/fsdbeta-megathread-for-all-fsd-beta-discussions.18878/page-89#post-322871

[YTV 6, 7:37–8:00 min]

> It seems to be predicting other vehicles' intentions as well and oftentimes seems like it's planning ahead for its maneuvers.
> [YTV 6, 8:50–9:34 min]

Because of the high uncertainty tolerance of the human drivers, the ADS in these Tesla cars were able to learn from experience, correct their own mistakes, and improve their performance over time:

> What an awesome update 10.12 ended up being. It includes new training data from over 250 000 real world video clips and an absolutely massive change log …It learns from actual experiences and improves itself just like we do.
> [YTV 1, 11:27–11:36 min]

In the next section, we discuss the significance of these findings in relation to the ways by which humans learn from uncertainty and how such learning impacts joint human-AI task performance. We link our discussion to more recent developments in generative AI and the implications of our analysis for better theorizing human-machine interdependencies in organizational contexts.

5. Discussion

As evident from our analysis, the joint task performance of humans and AI systems is rife with challenges in delegating, supervising and controlling task performance (Baird & Maruping, 2021). The four aggregate themes emerging from our analysis of Tesla's ADS reveal distinct variations of learning from uncertainty in real time and over time. These variations are shaped by humans' individual uncertainty tolerance that impact the joint human-AI task performance. We show that depending on their uncertainty preferences, humans may explore different modes of AI use including *uncontrolled automation*, *limited automation*, *expanded automation*, and *controlled automation*. In this section we present and discuss our theoretical model of these modes of AI use, including how humans move between these modes and how the cycle of learning from uncertainty feeds into new AI systems. Fig. 1 illustrates our model.

5.1. From uncontrolled to controlled automation: reducing uncertainty tolerance

The crash investigation reports provide evidence that tasks are not performed in a vacuum, rather, they are always part of broader complex systems and can generate cascading effects in those systems (e.g., crashing into other cars). Such cascading effects are very difficult to predict and can lead to multiple pathways because of the causal interdependencies between interacting humans and AI systems (Benbya, Nan, Tanriverdi, & Yoo, 2020). Humans with high uncertainty tolerance risk falling victims of such cascading effects. High uncertainty tolerance can subject humans to overconfidence in the capabilities of AI systems and automation complacency (Endsley, 2017; Parasuraman & Manzey, 2010). In such cases, humans can become completely disengaged and inattentive to the task letting AI systems run into unknown unknowns.

In high uncertainty tasks such as medical and criminal justice decision making and even financial trading, high uncertainty tolerance can generate unprecedented outcomes not only for individuals but also for the organizations within which they operate. For example, in high frequency trading, AI systems are delegated the task to automatically process millions of orders across financial instruments and on several trading platforms. While such automation is meant to be supervised by humans, AI systems are often left to autonomously respond to other agents placing buy or sell orders, because they can (computationally) handle the scale and speed of these orders better than humans (Van Kervel & Menkveld, 2019). The same scale and speed of these orders, however, may cause, what is called, a "flash crash", with AI systems incorrectly interpreting trends and engaging in mass order purchases or sales, leaving humans unable to react. Thus, humans' exploration of automation options by AI systems without monitoring them, can let AI systems engage in false exploitation of erroneous knowledge. Such automation can generate significant, negative consequences not only for individual tasks, but also for systems of tasks.

In these and similar organizational settings there are unintended effects from the high uncertainty tolerance of one actor for other actors. As others have noted, humans want to, but fail to, make good delegation decisions because of overconfidence in their own performance and in their abilities to learn from uncertainty (Fugener, Grahl, Gupta, & Ketter, 2021). Such overconfidence leads to a higher degree of inattentiveness (e.g., falling asleep on the wheel, as we saw in one of the crash investigation reports) that could lead to an inability to mitigate uncertainty when it emerges. Automation studies have found that out-of-the-loop performance leads to less attention to the task and the task context (Parasuraman & Manzey, 2010) and a loss of situation awareness (Endsley, 2017). This loss of situation awareness has negative effects on humans' ability to mitigate uncertainty, especially in high uncertainty tasks including driving, but also aviation (Johnson, Duda, Sheridan, & Oman, 2017).

It is only when there are automation failures that humans begin to re-evaluate their delegation decisions (Logg, Minson, & Moore, 2019), but such re-evaluation may in fact generate negative learning effects. Although, learning from automation failures can help improve the design and safety of AI systems, it will also generate issues of mistrust against AI systems. Mistrust can lead to low uncertainty tolerance in the future, even though it is often humans' automation complacency and overconfidence that lead to fatal outcomes the first place. Indeed, studies have shown that, in high uncertainty tasks that involve other humans, trust in humans is higher than in AI systems (see Glikson & Woolley, 2020 for a review). Exactly because trust is a moderator of technology acceptance,

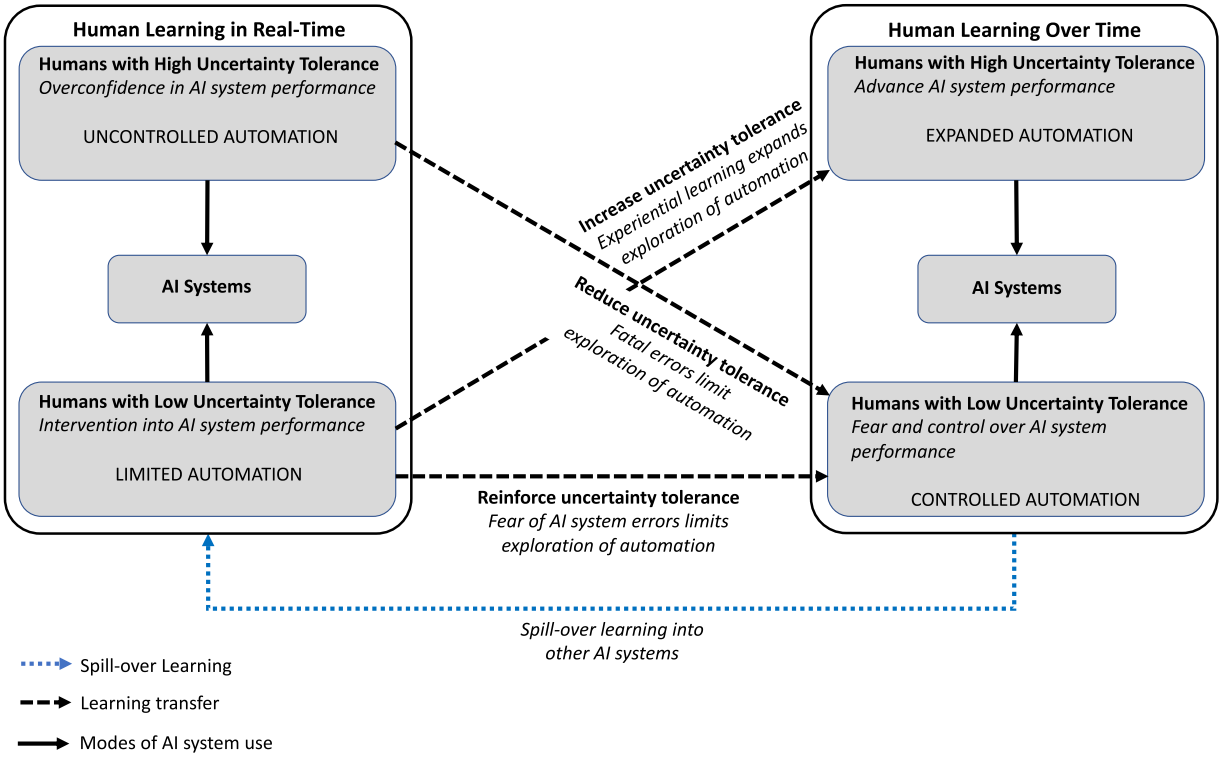

Fig. 1. A model of human learning in joint human-AI performance.

misuse of AI systems (i.e., an inappropriate or blind exploration of automation without appropriate exploitation of human agents' knowledge of the task in situ) can lead to disuse of AI systems over time (Glikson & Woolley, 2020). Fatal accidents and all the negative publicity that follows them generate fear over automation and a mistrust over AI systems' ability to perform a task without errors.[16] This has spurred a range of experimental studies seeking to improve how AI systems communicate explanations before and after a critical incident as a means of convincing the public and regulators of their automation capabilities and to promote trust in the technology (see Du et al., 2019).

5.2. *From limited to controlled automation: reinforcing low uncertainty tolerance*

Continuing from the above discussion, even humans with low uncertainty tolerance may become fearful of automation failures as their mistrust in AI system is reinforced. These humans are risk averse and will often be too critical of the AI system's performance. In this sense, it represents a missed opportunity. These humans will resort to exploiting existing knowledge and deactivating AI systems soon after activating them because they are defiant of even the slightest form of automation. Humans with low uncertainty tolerance cannot accept losing control over the task. As discussed earlier, publicised automation failures cause increased anxiety to human agents and algorithmic aversion (Dietvorst, Simmons, & Massey, 2018; Filiz et al., 2023). If such algorithmic aversion persists, there are risks of human agents losing confidence and trust over the capabilities of AI system, never activating them in joint task performance. "Low trust in highly capable technology would lead to disuse and high costs in terms of lost time and work efficiency" (Glikson & Woolley, 2020, p. 631).

Humans' confidence in AI systems lie in their ability to understand how these systems perform and why they perform in certain ways but not others (Hoff & Bashir, 2015). In building such confidence in AI system performance, humans also build confidence in their own performance, as the outcomes of their joint performance may align. For example, in healthcare, a correct diagnosis by an AI system may confirm the diagnosis of a human doctor, increasing the latter's confidence (Lebovitz et al., 2022). But before this can be achieved, humans need to understand how AI systems perform, especially when outcomes differ. The point here is not to build confidence around human-like performance; but rather to build confidence in an AI system's capabilities to mitigate uncertainty on their own and jointly with humans. We acknowledge that some tasks may not require AI systems – i.e., more simple interpretable models such as decision trees would suffice (Rudin, 2019). However, for tasks that involve estimating probabilities with multiple variables, and computing highly complex conditional correlations (Chen, Jiang, Li, Jia, & Ghamisi, 2016), AI systems will be required

[16] See for example this article in the BBC https://www.bbc.com/news/business-44159581

and will most often appear black-boxed to human agents.

Learning from uncertainty in high uncertainty tasks in the midst of inscrutable AI system performance, as in the case of 'phantom' breaking in our empirical analysis, requires an understanding of the role of humans outside the task performance (i.e., the designers). Collecting and labelling lessons learned from such inscrutable cases so as to feed those into future upgrades of the AI systems becomes critical in improving the joint task performance. However, as we observed in our analysis, such learning may never take place in time or in specific locations. Learning takes place if humans activate AI systems in specific scenarios (e.g., right turn in a busy intersection) and specific locations (e.g., Middle Tennessee). If such activations never take place, humans will continue to be faced with inscrutability, something that will hamper their confidence in AI systems, as discussed above.

Indeed, AI systems in some organizational settings may never benefit from the type of crowd learning we observed in the Tesla case. For example, the learning that takes place for AI systems used in individual hospitals will vary based on the number of patients, type or number of diagnostic and other devices that integrate with those systems, the number of experts available at that organization and the modalities contained within the data sets, all of which may generate false predictors (Constantinides & Fitzmaurice, 2018; Darzidehkalani, Ghasemi-Rad, & van Ooijen, 2022). Such learning may generate myopias to existing organizational knowledge (Balasubramanian, Ye, & Xu, 2021). All of these issues may generate low uncertainty tolerance with negative effects for joint human-AI task performance. In turn, AI systems will not be given the opportunity to learn to mitigate uncertainty on their own. Likewise, humans will never learn to mitigate uncertainty in joint with AI systems. In consequence, in these situations, there is no joint human-AI task performance.

5.3. From limited to expanded automation: increasing uncertainty tolerance

In contrast to the above, humans with low uncertainty tolerance but with full situational awareness of the task at hand will often allow AI systems to perform autonomously, intervening and correcting errors when needed. Our analysis shows that such practices emerge through learning from joint task performance. In other words, humans learn to mitigate uncertainty by themselves and in joint with AI systems by departing from their preconceived model of how a task *should* be performed to enable performance that is occasioned by, and responsive to, the capabilities and limitations of AI systems. Humans and AI systems can augment one another, all the while they continue to monitor each other's performance.

The significance of this insight transcends organizational settings by pointing to what others have called "engaged augmentation", where humans relate their knowledge of the task to the knowledge of AI systems (Lebovitz et al., 2022). We argue that such engaged augmentation can only come with experience that enables humans to increase their uncertainty tolerance, while allowing AI systems to make errors under supervision and to correct those errors on their own. When humans explore automation options by AI systems over time, AI systems learn to improve their performance and to mitigate the uncertainty they face on their own.

As our analysis of Tesla's ADS has shown, an AI system may experience uncertainty due to an edge case, that is, an abnormal case that is totally normal for a human driver. For example, a garbage truck stopping in the middle of the road, in the absence of any traffic signals, was an edge case for Tesla's ADS. Because this case was inscrutable to the AI system, it proceeded to perform the task in a very unconventional way, that is, turning into someone's private driveway in order to overtake the garbage truck. At that point, the AI system created uncertainty for the human driver, causing anxiety, lack of tolerance and intervention. In other words, the AI system's exploration of unknown options created the need for less exploration of automation by human drivers and more exploitation of existing knowledge. One could easily argue that more experience with such edge cases could have led to a better understanding of why the AI system performed the task as such, while also allowing the system to learn how a garbage truck performs. A counter argument would also raise the question of safety being compromised – one of the key objectives of Tesla's ADS – thus, necessitating a human intervention. A supervising human agent needs to be able to distinguish cases where exploration of automation options is possible from cases where more careful exploitation of existing knowledge (and intervention) is required. Exactly because this relationship is dynamic, however, there is no rule book to predetermine when one emerges over the other. In tasks where feedback on the performance is delayed, sparser and more ambiguous it is harder to predict which choice offers the highest returns while mitigating risks (Kahneman, Sibony, & Sunstein, 2021). As noted earlier, high uncertainty tasks generate higher interdependencies between humans and AI systems such that even a small error can rapidly escalate into further errors with potentially devastating outcomes for everyone involved (Perrow, 1999). In such high uncertainty tasks, humans will be faced by known unknowns and unknown unknowns simultaneously (Packard & Clark, 2020). It is up to humans to select where to direct their attention and engage in counterfactuals by means of evaluating uncertainty in their task performance (Tannenbaum, Fox, & Ülkümen, 2017).

Augmentation needs to be mutual between humans and AI systems, such that humans monitor and correct mistakes by AI systems and vice versa. Specifically, by observing AI systems autonomously perform a task, humans can also learn about new options with which to augment their existing knowledge. Recent developments around generative AI technologies exemplifies how humans can be augmented with new options that can improve their task performance. One study identified 18 different consulting tasks and ran experiments to assess whether consultants using ChatGPT4 could outperform those who did not (Dell'Acqua et al., 2023). In the majority of tasks (including analytical, creative and writing tasks) those who used generative AI performed significantly better than those who did not. The study also found that generative AI worked as a skill enhancer, with consultants experiencing a jump in their performance. However, the study also found that for some tasks that required a deeper understanding of the context, consultants working with generative AI performed worse than those who worked on their own. These consultants missed important information and were influenced by AI recommendations that were blind to this information. This study confirms our own findings that, in the context of edge cases, humans should not rely on automation but rather a careful delegation of the division of labor, allocating responsibilities to AI systems based on their strengths and capabilities, while humans retaining other responsibilities. The learning effects

of such delegation will only be observed over time, as they jointly learn to perform the task with AI systems, in situ.

Before such learning effects can be captured, however, tasks that are causally ambiguous and embedded in larger complex systems like driving or medical diagnostics, learning from uncertainty will be dependent on actively engaged, supervising humans. For example, in the context of driving, some have argued that the only way we can ever achieve full automation is to close off parts of the road infrastructure to enable cars to run on dedicated lanes as a rail system (Choe, Oettl, & Seamans, 2021) or to contain AI systems' task performance in prespecified areas, as in the case of Voyage (recently acquired by Cruise) offering rides to residents within gated, retirement communities.[17] In these cases, humans are remotely supervising the performance of AI systems, ready to intervene and even take over control in the event that AI systems are faced with edge cases. Once those edge cases become normal for AI systems, then they can begin to augment humans, monitoring their performance, alerting for errors and even intervening to correct human errors.

The key message of our theoretical model is that the more humans learn from uncertainty while using AI systems the more they can augment their own, as well as the AI systems' performance. The automation-augmentation frontier will continue to expand as technologies improve over time (Berente et al., 2021) and as humans use those technologies while learning from uncertainty. As a consequence, it is increasingly rare that humans will meet an AI system for the first time; our next encounter feeds off the experiences we already have accumulated with AI systems as captured by the feedback loop in our model (Fig. 1). In this continuous feedback loop, our uncertainty tolerance needs to be monitored and adjusted according to an understanding of our joint task performance with AI systems. We should not "fall asleep at the wheel" by becoming overconfident of AI system performance (Dell'Acqua, 2022). We should instead remain active in the loop, allocating task responsibilities based on our strengths and capabilities and those of emergent AI systems.

6. Practical implications

In conclusion, we discuss implications for the design of AI systems, policy into delegation in joint task performance, as well as use of data to improve learning from uncertainty.

First, the coming of age of generative AI systems such as ChatGPT-4 by Open AI, LLaMa by Meta and Bard by Google have created anxiety and protest against what constitutes the best and most responsible approach to design AI systems.[18] Such discussions could lead to the false assumption that learning from uncertainty in the joint task performance of human and AI systems is very much implicated in the design of these systems. For example, Tesla has been criticized that by relying solely on computer vision sensors they limit the identification of some objects and some dynamic trajectories in the environment, unlike other cars like Alphabet's Waymo that use a fusion of sensors, including LIDAR, cameras, and ultra-sonic sensors. These distinct hardware configurations are also complemented with unique neural network architectures and machine learning models to evaluate emergent uncertainty in the driving task. One could argue that if all cars used the same – 'best of' – hardware and software configurations, then we would be better equipped to learn from uncertainty. One could also argue that if all cars adopted the same – 'best of' – learning processes, such that all cars became part of a larger 'intelligent collective', each learning from the edge cases experienced by one another, we would also be in a better position to learn from uncertainty. However, just like there are different types of human drivers with varied levels of uncertainty tolerance, there are different types of AI systems produced by different manufacturers, all of which perform the driving task with varied learning algorithms, each with unique benefits and limitations. It is exactly such heterogeneity – not "monocropping" (Scott, 1998), i.e., the imposition of blueprints based on idealized versions of big tech – that has been known to contribute to a diversity of learning experiences, competition, innovation and eventually improved task performance. Indeed, there is a strong argument that, developing open-sourced, crowd-based generative AI systems is the only way to resist the homogeneity of dominant technologies such as ChatGPT that may generate all sorts of biases, inequalities and harm.[19]

As such, we are critical of calls for standardization such as the Society of Automotive Engineers (SAE) International Standard (SAE, 2014) in car automation. Influenced by the ways car manufacturers and technology providers differentiate between evolutionary levels of car automation by means of technology configurations, the SAE standard relies on the assumption that ADS evolve by design. This is why the SAE presents a linear evolution of car automation that does not consider variations in the joint task performance of human and AI systems. We argue that designers of AI systems should not converge on a single standard, nor conform to homogeneous socio-technical configurations but rather embrace (and encourage) their diversity and heterogeneity, which make up the pre-condition for a more robust strategy. Designers should, however, be explicit about the limitations in the task performance of their AI systems vis-à-vis other agents. This would only help increase the safe use of AI systems across different tasks as it will also help humans to learn from uncertainty with better long-term outcomes.

Second, although much has been discussed about the inscrutability and black-box nature of AI systems and the challenges this poses for the learning from uncertainty by humans, we show that it takes two to tango. Joint task performance is *brittle,* never perfect, but always subject to errors caused by both humans and AI systems and their relationality. Augmentation and automation are not separate processes (Raisch & Krakowski, 2021) nor are they 'human-free' processes; automation should not mean that humans are out-of-the-loop despite the expectation that this should be the case. AI systems should not be expected to autonomously take all the weight of task performance, especially in high uncertainty tasks, such as driving and medical diagnostics. Rather humans are required to be attentive,

[17] See https://news.voyage.auto/voyages-first-self-driving-car-deployment-29c7688c6a1

[18] A number of prominent artificial intelligence experts, tech entrepreneurs and scientists called on *"all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4"* in an open letter https://futureoflife.org/open-letter/pause-giant-ai-experiments/

[19] See https://techcrunch.com/2023/03/14/meet-the-team-developing-an-open-source-chatgpt-alternative/

monitoring the task performance of AI systems and intervening when needed just like AI systems are expected to do the same. Joint task performance involves active delegation between agents to learn from uncertainty. Delegation need not be collapsed to one-to-one agent interactions and mode transitions on tasks or subtasks. Delegation can be scaled to groups and larger collectives that can augment joint task performance. Deciding how such delegation takes place (which agent or collective of agents performs which subtask, which agent supervises and how they intervene in the face of uncertainty) becomes important.

In particular, collective human-AI task performance needs to incorporate both design considerations of delegation via modularization of system components, as well as policy considerations for governing learning between diverse types of agents. This is already becoming important as organizations begin to deploy multimodal generative AI technologies for different functions. Each function requires domain knowledge from human experts that can be fed into modular AI system components via text, image and other types of data to perform functional tasks. Collectively, these functions produce knowledge that can inform organizational decisions and, thus, need to be fused together. Just like generative AI technologies are architected across different layers and modular components to learn to perform a task,[20] so is human learning within organizations and their broader ecosystems. However, fusing the two in ways that collectives of human and AI systems can learn from uncertainty and jointly perform a task is far from trivial and would require close collaboration between designers, policy makers and regulatory agencies.

The final implication concerns the use of data to improve the learning from uncertainty in joint task performance. Although there is much discussion on the need to get rid of 'biases' in the data that are intentional and historically persistent and that are consequential to our practices, the reality is that there is no such thing as unbiased data. Data are collected, labelled and given weight based on the individual experiences and preferences of humans, including both end users and designers. AI systems inherit those preferences in their learning models, so there is a non-virtuous recycling of biases in the data. As others have noted, large datasets "overrepresent hegemonic viewpoints and encode biases potentially damaging to marginalized populations" (Bender, Gebru, McMillan-Major, & Shmitchell, 2021, p. 610). We can only mitigate the risks of such harm, by "creating datasets as large as can be sufficiently documented" (ibid. 610), evaluating the cases for which those data are to be used, while also being sensitive to the values of participant stakeholders who may be affected by the joint tasks performance of human and AI systems. Field experiments in contained task environments could test the efficacy of such an approach and the impact it has on the learning from uncertainty in joint task performance. Lessons learned from such experiments could then be scaled up to larger collectives. Our efforts to design AI systems that can automate and augment tasks previously performed by humans alone need to take into account how humans learn from uncertainty, including their uncertainty tolerance, to help them move between different modes of AI use with more confidence.

Author statement

We would like to submit the attached manuscript, "Human-AI Joint Task Performance: Learning from Uncertainty in Autonomous Driving Systems," for consideration for possible publication at *Information & Organization.*

This paper (or closely related research) has not been published or accepted for publication. It is not under consideration at another journal or at *Information & Organization.*

CRediT authorship contribution statement

Panos Constantinides: Conceptualization, Data curation, Formal analysis, Methodology, Writing – original draft, Writing – review & editing. **Eric Monteiro:** Conceptualization, Writing – original draft, Writing – review & editing. **Lars Mathiassen:** Conceptualization, Writing – original draft, Writing – review & editing.

References

Baird, A., & Maruping, L. M. (2021). The next generation of research on IS use: A theoretical framework of delegation to and from agentic IS artifacts. *MIS Quarterly, 45*(1), 315–341.

Balasubramanian, N., Ye, Y., & Xu, M. (2021). Substituting human decision-making with machine learning: Implications for organizational learning. *Academy of Management Review, 47*(3), 448–465.

Banks, V. A., Eriksson, A., O'Donoghue, J., & Stanton, N. A. (2018). Is partially automated driving a bad idea? Observations from an on-road study. *Applied Ergonomics, 68*, 138–145.

Banks, V. A., Stanton, N. A., & Harvey, C. (2014). Sub-systems on the road to vehicle automation: Hands and feet free but not 'mind' free driving. *Safety Science, 62*, 505–514.

Bauchwitz, B., & Cummings, M. (2020). *Evaluating the reliability of Tesla model 3 driver assist functions*. Collaborative Sciences Center for Road Safety: Duke University.

Benbya, H., Nan, N., Tanriverdi, H., & Yoo, Y. (2020). Complexity and information systems research in the emerging digital world. *MIS Quarterly, 44*(1), 1–17.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610–623).

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly, 45*(3).

Borowsky, A., Shinar, D., & Oron-Gilad, T. (2010). Age, skill, and hazard perception in driving. *Accident Analysis & Prevention, 42*(4), 1240–1249.

Burton, J. W., Stein, M. K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making, 33*(2), 220–239.

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research, 56*(5), 809–825.

[20] See for example AWS's Generative AI Stack https://aws.amazon.com/blogs/machine-learning/welcome-to-a-new-era-of-building-in-the-cloud-with-generative-ai-on-aws/

Chen, Y., Jiang, H., Li, C., Jia, X., & Ghamisi, P. (2016). Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing, 54*(10), 6232–6251.
Choe, D., Oettl, A., & Seamans, R. (2021). What's driving entrepreneurship and innovation in the transport sector?. In *NBER working paper 27284*.
Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 362*(1481), 933–942.
Constantinides, P., & Fitzmaurice, D. A. (2018). Artificial intelligence in cardiology: Applications, benefits and challenges. *British Journal of Cardiology, 25*(3), 86–87.
Darzidehkalani, E., Ghasemi-Rad, M., & van Ooijen, P. M. A. (2022). Federated learning in medical imaging: Part I: Toward multicentral health care ecosystems. *Journal of the American College of Radiology, 19*(8), 969–974.
Dell'Acqua, F. (2022). *Falling asleep at the wheel: Human/AI collaboration in a field experiment on HR recruiters* (Working paper).
Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., … Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. In *Harvard Business School Technology & Operations Mgt. Unit Working Paper* (24–013).
Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126.
Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science, 64*(3), 1155–1170.
Du, N., Haspiel, J., Zhang, Q., Tilbury, D., Pradhan, A. K., Yang, X. J., & Robert, L. P., Jr. (2019). Look who's talking now: Implications of AV's explanations on driver's trust, AV preference, anxiety and mental workload. *Transportation Research part C: Emerging Technologies, 104*, 428–442.
Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint*. arXiv:2303.10130.
Endsley, M. R. (2017). From here to autonomy: Lessons learned from human–automation research. *Human Factors, 59*(1), 5–27.
Endsley, M. R., & Kaber, D. B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics, 42*(3), 462–492.
Eriksson, A., Banks, V. A., & Stanton, N. A. (2017). Transition to manual: Comparing simulator with on-road control transitions. *Accident Analysis & Prevention, 102*, 227–234.
Eriksson, A., & Stanton, N. A. (2017). Takeover time in highly automated vehicles: Noncritical transitions to and from manual control. *Human Factors, 59*(4), 689–705.
Favarò, F. M., Nader, N., Eurich, S. O., Tripp, M., & Varadaraju, N. (2017). Examining accident reports involving autonomous vehicles in California. *PLoS One, 12*(9), Article e0184952.
Felten, E. W., Raj, M., & Seamans, R. (2023). *Occupational heterogeneity in exposure to generative ai*. Available at SSRN 4414065.
Filiz, I., et al. (2023). The extent of algorithm aversion in decision-making situations with varying gravity. *PLoS One, 18*(2).
Fugener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI. *MIS Quarterly, 45*(3).
Gavetti, G., & Levinthal, D. (2000). Looking forward and looking backward: Cognitive and experiential search. *Administrative Science Quarterly, 45*(1), 113–137.
Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition, 173*, 34–42.
Gkeredakis, M., & Constantinides, P. (2019). Phenomenon-based problematization: Coordinating in the digital era. *Information and Organization, 29*(3), 100254.
Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals, 14*(2), 627–660.
Gottlieb, J., & Oudeyer, P.-Y. (2018). Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience, 19*, 758–770.
Griffin, M. A., & Grote, G. (2020). When is more uncertainty better? A model of uncertainty regulation and effectiveness. *Academy of Management Review, 45*(4), 745–765.
Hartmann, K., & Wenzelburger, G. (2021). Uncertainty, risk and the use of algorithms in policy decisions: A case study on criminal justice in the USA. *Policy Sciences, 54*(2), 269–287.
Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407–434.
Jain, H., Padmanabhan, B., Pavlou, P. A., & Raghu, T. S. (2021). Editorial for the special section on humans, algorithms, and augmented intelligence: The future of work, organizations, and society. *Information Systems Research, 32*(3), 675–687.
Johnson, A. W., Duda, K. R., Sheridan, T. B., & Oman, C. M. (2017). A closed-loop model of operator visual attention, situation awareness, and performance across automation mode transitions. *Human Factors, 59*(2), 229–241.
Jussupow, E., Spohrer, K., Heinzl, A., & Gawlitza, J. (2021). Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. *Information Systems Research, 32*(3), 713–735.
Kahneman, D., Slovic, S. P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.
Kahneman, Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgement*. New York: Little, Brown Spark.
Lebovitz, S., Lifshitz-Assaf, H., & Levina, N. (2022). To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organization Science, 33*(1), 126–148.
Liu, Q., Wang, X., Wu, X., Glaser, Y., & He, L. (2021). Crash comparison of autonomous and conventional vehicles using pre-crash scenario typology. *Accident Analysis & Prevention, 159*, Article 106281.
Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes, 151*, 90–103.
March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science, 2*(1), 71–87.
Monteiro, E., Constantinides, P., Scott, S., Shaikh, M., & Burton-Jones, A. (2022). Editor's comments: Qualitative methods in IS research: A call for phenomenon-focused problematization. *MIS Quarterly, 46*(4), iii–xix.
Mousavi, S., & Gigerenzer, G. (2014). Risk, uncertainty, and heuristics. *Journal of Business Research, 67*(8), 1671–1678.
Nilsson, J., Strand, N., Falcone, P., & Vinter, J. (2013). Driver performance in the presence of adaptive cruise control related failures: Implications for safety analysis and fault tolerance. In *Proceedings of the 2013 IEEE/IFIP 43rd international conference on dependable systems and networks workshops (DSN-W9)*.
Packard, M. D., & Clark, B. B. (2020). On the mitigability of uncertainty and the choice between predictive and nonpredictive strategy. *Academy of Management Review, 45*(4), 766–786.
Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors, 52*(3), 381–410.
Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 30*(3), 286–297.
Perrow, C. (1999). *Normal accidents: Living with high-risk technologies* (2nd ed.). Princeton, NJ: Princeton University Press.
Rai, A., Constantinides, P., & Sarker, S. (2019). Editor's comments: Next-generation digital platforms: Toward human–AI hybrids. *MIS Quarterly, 43*(1), iii–x.
Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. *Academy of Management Review, 46*(1), 192–210.
Reich, T., Kaju, A., & Maglio, S. J. (2022). How to overcome algorithm aversion: Learning from mistakes. *Journal of Consumer Psychology, 33*(2), 285–302.
Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215.
SAE. (2014). Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems. Standard J3016_201401, Issued 16 January 2016. Available at: https://www.sae.org/standards/content/j3016_201401/.
Sagberg, F., & Bjørnskau, T. (2006). Hazard perception and driving experience among novice drivers. *Accident Analysis & Prevention, 38*(2), 407–414.
Samuel, S., Borowsky, A., Zilberstein, S., & Fisher, D. L. (2016). Minimum time to situation awareness in scenarios involving transfer of control from an automated driving suite. *Transportation Research Record, 2602*(1), 115–120.
Scott, J. C. (1998). *Seeing like a state: How certain schemes to improve the human condition have failed*. Yale University Press.
Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics, 69*, 99–118.

Tannenbaum, D., Fox, C. R., & Ülkümen, G. (2017). Judgment extremity and accuracy under epistemic vs. aleatory uncertainty. *Management Science, 63*(2), 497–518.
Van Kervel, V., & Menkveld, A. J. (2019). High-frequency trading around large institutional orders. *The Journal of Finance, 74*(3), 1091–1137.
Walker, A. R., Luque, D., Le Pelley, M. E., & Beesley, T. (2019). The role of uncertainty in attentional and choice exploration. *Psychonomic Bulletin & Review, 26*, 1911–1916.