INL/EXT-20-59469

Light-Water Reactor Sustainability Program

Towards a Deeper Understanding of Automation Transparency in the Operation of Nuclear Plants

August 2020

U.S. Department of Energy

Office of Nuclear Energy

INL/EXT-20-59469

Towards a Deeper Understanding of Automation Transparency in the Operation of Nuclear Plants

Gyrd Skraaning, Jr., Greg A. Jamieson, and Jeffrey Joe

August 2020

Prepared for the
U.S. Department of Energy
Office of Nuclear Energy

ABSTRACT

Automation technologies have the potential to reduce operations and maintenance costs, ensure reliable power generation and safety, and thereby contribute to extending the lifespan of nuclear power plants. To fulfill this aspiration, automation behaviors have to be understandable and predictable to human operators and traceable for license holders and regulators in the event of automation failures. The current report investigates how automation transparency as a systems design principle can keep operators sufficiently informed about the inner workings of automation or make fail-safe automation run invisibly in the background to free up operator capacity.

FOREWORD

The present contract report arose from discussions between the Institutt for energiteknikk and Idaho National Laboratory following the presentation of HWR 1250 at the May 2019 Extended Halden Programme Group Meeting in Sandefjord, Norway. In this foreword, we review the content of HWR-1250.

The first half of HWR-1250 draws a distinction between seeing-through and seeing-into automation transparency. The human factors discipline adopts the seeing-into perspective: the principle that the responsibilities, capabilities, goals, activities, inner workings, performance, and effects of automation should be observable/visible to the operator1. This perspective on transparency, however, is antithetical to the perspective adopted in other fields. In teleoperation, computer network design, and mobile communication infrastructure for example, automation transparency refers to the experience of direct interaction with a task through a technology medium so well designed as to appear invisible. This alternative use has been referred to as seeing-through automation transparency. The work report makes the argument that the opposing semantic meanings of these two uses of automation transparency present a concern. The report advocates for the adoption of the alternative phrase, automation apparency2 to refer to the seeing-into connotation of transparency3.

HWR-1250 proceeds with a brief description of the Situation Awareness-Based Agent Transparency (SAT) model, introduced by Chen and colleagues to facilitate human collaboration with smart and independent computer agents in the military command and control domain. Although SAT has dominated the human factors literature on automation transparency (i.e. transparency in the seeing-into sense) for more than 5 years, the HWR-1250 offers critiques of the model's weak conceptual foundations, ambiguity in content and structure, conceptual overreach, and unconvincing early empirical validation. We expand on the empirical evidence in this report.

In a final section of the work report, we explore how the notion of transparency arises in human-robot interaction (HRI). After recounting a case study, we discuss the central concern about human trust in the robot and the misleading presumption that automation transparency (in the seeing-into sense) emerged as a response to the trust problem. While these two notions are often intertwined, trust has many bases that are unrelated to information disclosure; and transparency as a design principle has many anticipated effects beyond trust calibration. That said, exploring the HRI literature on transparency gave us confidence that further insight could be found by exploring other literatures; an approach we follow up with in this report.

1 This definition is updated from the HWR-1250 as we continue to encounter transparency content expectations in the literature. We include this footnote explicitly as we have previously noted how inconsistent use of technical terms in human factors has led to confusion.

2 Credit to Sheridan and Verplank (1978) for introducing this term.

3 Throughout most of this report we continue to use "seeing-into transparency", although we revert to the more concise and descriptive "apparency" when we present a taxonomy of automation transparency design frameworks.

CONTENTS

FIGURES

TABLES

ACRONYMS

AFRL	Air Force Research Laboratory
AI	artificial intelligence
AOA	angle of attack
B737 MAX	Boeing 737 MAX
CBP	computer-based procedures
COSS	Computerized Operator Support Systems
DARPA	Defense Advanced Research Projects Agency
DOE	Department of Energy
FAA	Federal Aviation Administration
HFE	human factors engineering
HRI	Human-robot interaction
I&C	instrumentation and control
IEEE	Institute of Electrical and Electronics Engineers
IFE	Institutt for energiteknik
INL	Idaho National Laboratory
IoT	internet of things
LWRS	Light-Water Reactor Sustainability
MCAS	Maneuvering Characteristics Augmentation System
ML	machine learning
NPPs	nuclear power plants
O&M	operations and maintenance
R&D	research and development
SAT	Situation Awareness-Based Agent Transparency
U.S.	United States
XAI	explainable artificial intelligence

Towards a Deeper Understanding of Automation Transparency in the Operation of Nuclear Plants

1. MOTIVATION AND OBJECTIVE

In the United States (U.S.), commercial nuclear power plants (NPPs) generate approximately one-fifth of the reliable baseload electricity that powers the nation's economy. Because of this important function, owners and operators are actively working to not only maintain but continuously improve and extend the operating life of existing NPPs. The Light-Water Reactor Sustainability (LWRS) Program, which is sponsored by the U.S. Department of Energy (DOE) Office of Nuclear Energy, has the mission to perform research and development (R&D) that establishes the technical bases for NPP life extension. One research area in the LWRS Program is the Plant Modernization Pathway, which includes human factors R&D, human factors engineering (HFE), and ergonomics. LWRS Program researchers in this pathway conduct targeted R&D to address challenges with the legacy instrumentation and control (I&C) systems by helping to design, demonstrate, and deploy digital I&C technologies. In doing this, the LWRS Program not only helps ensure legacy I&C systems are not a life-limiting factor for U.S. commercial NPPs but also helps enable full plant modernization through broad innovation and digitalization, thereby allowing the industry to develop advanced concepts of operations that improve the business case for the continued operation of NPPs.

Automation technologies hold promise to make NPPs cost-competitive with other forms of electrical power generation. Currently, for many commercial U.S. NPPs, the most significant cost associated with operating the plant is in operations and maintenance (O&M). Automation is therefore seen as a way to reduce O&M costs while maintaining exceptionally high levels of safety. Two examples from the LWRS Program are Computerized Operator Support Systems (COSS) and computer-based procedures (CBP).

LWRS Program researchers at Idaho National Laboratory (INL) have been developing multiple versions of a COSS to assist operators in plant monitoring, fault diagnosis, and fault mitigation[4]. The COSS does not perform operator actions, but rather performs rapid assessments, computations, and provides recommendations to the operators. These automated information analysis and decision selection functions seek to reduce crew workload and augment operator judgment and decision-making during fast-moving, complex events (Thomas, Boring, Lew, Ulrich, and Vilim, 2013). As COSS are considered for implementation in NPP operations, a key question will be the extent to which operators will need to understand their underlying control logic and behavior.

Another example of NPP automation is in the development of CBP and other dynamic instruction systems (e.g., Oxstrand and Le Blanc, 2012). LWRS Program researchers have been automating traditional, static, paper-based procedures. CBP functions include automatic place keeping, correct component verification, calculations, integration with soft controls, and selective enabling of procedure steps that are relevant to the operating context. Automatic execution of these functions should minimize administrative and operational errors; however, questions remain regarding what the operators need to know with respect to how these dynamic instructions work (Jamieson and Skraaning, 2019; Skraaning and Jamieson, 2020).

Two central tenets of safe operation of highly automated facilities are comprehensibility and traceability. Comprehensibility dictates that—for operators to rely appropriately on automation—its behavior should be understandable and predictable. Traceability dictates that—in the event of automation failures—license holders and regulators must understand how and why failures occurred to prevent recurrence of similar events. If U.S. NPPs are to become more highly automated, how can technology developers provide the necessary level of comfort and confidence among regulators and utilities to

4 More recent research seeks to apply COSS to support cyber diagnostics.

embrace future automation technologies? That is, how can technology developers make the automation comprehensible and traceable—particularly when the nature of the technology itself challenges these design objectives?

There are many important questions about the underlying design philosophy of automation that need to be understood before automation technologies can be more broadly deployed. One question gaining prominence is the degree to which operators should be informed about the inner workings of automation (i.e. by introducing transparency in the seeing-into sense). As noted in the foreword, this is a topic of ongoing interest within the Halden Reactor Project's Human Technology Organization[5] program. This report advances that investigation by 1) providing an overview of related Institutt for energiteknikk (IFE) and University of Toronto research activities, 2) summarizing empirical efforts to verify and validate the predominant automation transparency design framework, 3) recounting an industrial automation transparency design failure that led to significant loss of life and economic impacts, 4) expanding the human factors understanding of automation transparency as it manifests in artificial intelligence (AI), 5) introducing a taxonomy of automation transparency design approaches, and finally 6) offering preliminary insights on the implications of automation transparency design decisions in the context of future commercial NPPs.

2. UPDATE ON TRANSPARENCY-RELATED R&D ACTIVITY

2.1 Recent Activities, Progress and Developments

In an overlapping research program, we have gathered a group of loosely affiliated researchers around the topic of automation transparency (seeing-into). The group is led by Gyrd Skraaning (IFE) and Greg A. Jamieson (University of Toronto) and includes several University of Toronto graduate students (Rajabiyazdi, Quispe, Farooqi, and Gentile). This group has made several contributions since the publication of HWR-1250:

- IEEE P7001 Working Group – Transparency of Autonomous Systems (Rajabiyazdi – observer).

  The IEEE Standards Association initiated a Global Initiative on the Ethics of Autonomous and Intelligent Systems in April 2016. One of the five principles considered by the Committee is transparency. The Committee proposed the question: "How can we ensure that autonomous and intelligent systems are transparent?" To address this question, the IEEE Standards Association initiated IEEE P7001, a standard on the Transparency of the Autonomous System.

  P7001 is to provide guidance on how to develop autonomous technologies that represent why an autonomous system made a particular decision given the situation. The P7001 Working Group defined five groups of stakeholders including users, safety certifiers or agencies, accident investigators, lawyers or expert witnesses, and the wider public. As stated by Winfield (2019), the P7001 Chair, "For each of these stakeholder groups, P7001 is setting out measurable, testable levels of transparency so that autonomous systems can be objectively assessed and levels of compliance determined, in a range that defines minimum levels up to the highest achievable standards of transparency... P7001 will provide system designers with a toolkit for self-assessing transparency, and recommendations for how to achieve greater transparency and explainability." (p. 47–48).
- IEEE Access journal article submission – "A Review of Transparency in Human-Automation Interaction"—under revision (Rajabiyazdi, Jamieson, Skraaning, Mirjalali, Barnes, and Kinnear). Reviews received April 10, 2020.

5 Man Technology Organization through 2020.

- IEEE-SMC 2020 conference paper submission – "A Review of Transparency (seeing-into) Models"—in review (Rajabiyazdi, Jamieson and Skraaning).
- IEEE-SMC 2020 conference paper submission – "A Machine Learning-Based Micro-World Platform for Condition-Based Maintenance" (Quispe, Rajabiyazdi, and Jamieson). This open source platform will serve as the apparatus for subsequent seeing-into transparency studies.
- Skraaning Jr., G. and Jamieson, G. A. (2019). "Human Performance Benefits of The Automation Transparency Design Principle: Validation and Variation" Human Factors, https://doi.org/10.1177/0018720819887252.
- Early stages of a statistical meta-analysis of empirical transparency results (Rajabiyazdi). This project will yield an open source database of experiments.
- Initial investigation of transparency in the context of Automated Money Laundering detection (Farooqi).
- New project w/ Ericsson on Transparency in Condition-Based Maintenance in Industrial Settings (Gentile).

2.2 Updates on the Validation of SAT

In addition to the above activities, we have continued to monitor the ongoing efforts to validate the Situation Awareness-based Automation Transparency model. HWR-1250 provided an initial assessment of these efforts and progress since that report has not been exceptional. However, the SAT model is the most ambitious formulation of seeing-into transparency. The U.S. Air Force Research Laboratory (AFRL) has invested considerable effort in developing and validating this model. The model and its empirical foundations are more or less indicative of what we observe in more general seeing-into transparency research. We also interpret that this research program has concluded and that other research institutions may be picking up on this work. For all of these reasons, an update on SAT validation efforts seems prudent.

Table 1 summarizes the SAT validation publications that we have identified to date.

Table 1. Publications reporting or summarizing SAT validation efforts.

Source	Contents
Wright et al. (2015)	Two experiments on the effects of level of transparency on human performance in a military supervisory control task.
Wright et al. (2017)	An AFRL technical report; likely the forthcoming article noted above and covering three peer-reviewed articles (Wright et al., 2016a, 2016b, 2017).
Chen et al. (2018)	A summary article, including several peer-reviewed articles (Mercado et al., 2016; Selkowitz et al., 2016; Stowers et al., 2016) and one technical report (Wright et al., forthcoming).
Guznov et al. (2020)	One experiment on level of robot communication transparency on human performance in a supervised robot path following task.
Bhaskara et al. (2020)	Summarizes SAT validation findings to date.

Given the origin of the SAT model in a U.S. Department of Defence Research Laboratory, it is not surprising that SAT applications to date have been localized to automated decision aids in the military context. Despite this common content, however, the applications reflect a variety of experimental settings and tasks. This variety is accompanied by nontrivial inconsistencies in the design expression of the SAT.

That is, the contents of each design instance appear to differ. These differences in settings, tasks and model expressions limit the informativeness of the empirical results (i.e., the SAT model validation is inhibited by design verification).

Table 2 summarizes the findings to date by predicted human performance impact of increased transparency according to the SAT model.

Table 2. Summary of empirical results of SAT validation efforts to date.

Outcome	Findings
Task performance	Results sometimes as predicted (e.g., Wright et al., 2015) but have also been observed to fail to follow predicted trends across SAT levels (e.g., Wright et al., 2015; 2016).
Workload	Results statistically insignificant and weak (e.g., Selkowitz et al., 2016; Wright et al., 2016; Guznov et al., 2020) or significant and large for only a subset of workload dimensions (e.g., Guznov et al., 2020).
Trust	Results are at times compelling (e.g., Selkowitz et al., 2016) but at other times narrowly restricted to specific blocks of data (Mercado et al., 2016).
Situation Awareness	• No observations reported by Mercado et al. (2016), Stowers et al. (2016), or Wright et al. (2015). • Selkowitz et al. (2016) report several results corresponding with SAT model predictions, but the effect sizes are small (η_2 =0.6–0.7). Other predictions are not supported with statistically significant findings. • Guznov et al. (2020) observed no effect of transparency.

Given the results summarized in Table 2, it would be premature to accept SAT as either verified or validated as a transparency design framework. In terms of verification, both Bhaskara et al. (2020) and we have independently noted the inconsistencies in the expression of the model across several interface implementations for agent-based automation. These include inconsistencies within individual designs, where contents of communication specified as belonging to one level of the model are encoded in the interface at different (or redundant) levels. As well, inconsistencies can be observed between the interfaces designed by researchers working in the research facility that advanced the SAT model itself. These discordant expressions of the SAT model in interfaces intended to serve as research platforms impair verification efforts.

In terms of validation, the evidence supporting the predictive power of the SAT model is not compelling. Stronger evidence is to be found in relative improvements in task performance and trust across increasing (and cumulative) levels of the SAT model. However, the evidence remains fickle. The predictive power of the model would be more thoroughly assessed through the inclusion of a baseline control condition. Workload results have followed expectations in only narrow selections of available data. Situation awareness results are sparse and, where statistically significant, indicative of weak effects. These inconsistent empirical results cannot be considered a sufficient technical basis for validation of the SAT model.

3. THE BOEING 737 MAX 8 ACCIDENTS SEEN FROM AN AUTOMATION TRANSPARENCY PERSPECTIVE

The nuclear industry has a well-established practice of learning from the operating experiences of other safety-critical industries. In particular, we have the opportunity to gain insight into early implementations of, and experiences of operators working with, advanced automation. The Boeing 737 MAX accidents and the subsequent investigations and U.S. Congressional hearings provide a recent,

high-profile, and well-documented case that speaks directly to the seeing-into and seeing-through approaches to automation transparency. In doing so, it highlights an important issue that the nuclear industry will contend with in the near future. In this section, we explain the aircraft design features and flight control characteristics that have contributed to two well-publicized Boeing 737 MAX accidents (KNKT, 2019; AAIB, 2019; and Endsley, 2019).

3.1 Design of a See-through Automated System[6]

The Boeing 737 MAX was developed to compete with the Airbus A320neo (new engine option). It is extremely expensive to develop new airframes, so Boeing sought to modify their popular B737 design. However, the B737 MAX needed larger fuel-efficient engines that did not fit under the wings of the original B737 airframe. Larger engines were therefore placed slightly forward and higher up on the wing of the new model. The narrow-body fuselage design from 1968 remained unchanged.

Given the considerable market penetration of the B737 and the high cost of training pilots on new aircraft, the B737 MAX designers sought to replicate the flight control characteristics of the familiar B737. However, the new engine placement resulted in different flying characteristics that would make it more likely for pilots to angle the aircraft too steeply upwards during takeoff, risking a stall. Boeing therefore developed and installed flight control software that made flying the B737 MAX as similar as possible to previous B737 models, despite the differences in engine placement.

The Maneuvering Characteristics Augmentation System (MCAS) works silently in the background by automatically pushing the nose of the aircraft downward (see Figure 1). MCAS activates under manual flight when an angle of attack (AOA) sensor indicates that the climb angle is too elevated. This correction is achieved through manipulation of the horizontal stabilizers at the tail of the airplane and is referred to as an automatic stabilizer trim. The stabilizer manipulation was controlled by software that activated without pilots being aware of the intervention.

6 This subsection draws from a similar description in an upcoming technical report for the OECD Halden Reactor Project.

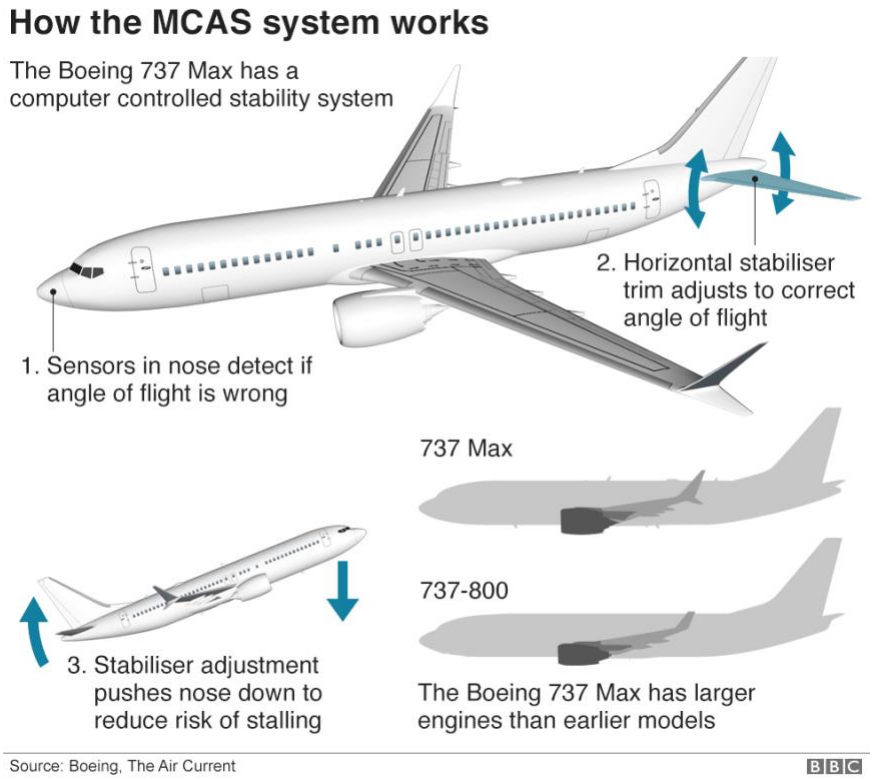

Figure 1. The MCAS automatic function (figure by The Air Current, BBC).

Two fatal Boeing 737 MAX 8 accidents (Lion Air flight 610 and Ethiopian Airlines flight 302) have been triggered by malfunctioning AOA sensors that fed incorrect information about the climb angle to MCAS. These instrumentation errors falsely activated MCAS and wrongly pushed the nose of the aircraft down towards the ground. In such cases, it is up to the pilots to intervene by manually trimming the stabilizer, but MCAS had incorrectly been presumed to be a fail-safe function that would operate reliably in the background. It was therefore implemented as black box automation with no way for pilots to infer the intentions, activation, and behavior of the system. MCAS was excluded from the transition training for the B737 MAX models, and the flight crew operations manual made no mention of the system. In both accidents, pilots ended up in a battle with a hidden automatic system that they could not understand or collaborate with.

3.2 Interpretation of B737 MAX Accidents from an Automation Transparency Perspective

The B737 MAX case study offers a revealing look into the seeing-through versus seeing-into automation transparency design approaches. MCAS worked according to the seeing-through principle. It was incorrectly presumed to be a fully automatic and fail-safe function that would operate in the background and never require human intervention. When faulty sensor information caused MCAS to behave erratically, pilots were unexpectedly thrust into a situation where automation transparency in the seeing-into sense could have helped them tremendously. They ended up fighting with silent black box automation that was virtually impossible to understand and control.

How did Boeing come to adopt the seeing-through approach to the design of the MCAS? As explained in the U.S. Congressional hearing on the B737 MAX accidents (2019), early design documentation reveals that Boeing had considered installing an indicator light that would have alerted

pilots to an MCAS failure (although not a mere triggering of the system). However, that indication was later integrated with a failure indicator for the speed trim systems of which Boeing came to consider MCAS a component.

Further evidence of Boeing's adoption of the seeing-through design approach can be found in email correspondence explicitly calling for the removal of MCAS from the flight crew operations manual and pilot training material (see Figure 2 and the U.S. Congressional hearing on the B737 MAX accidents [2019]). Boeing (with the approval of the Federal Aviation Administration [FAA]) deliberately transitioned from a seeing-into to a seeing-through automation transparency design approach. By some accounts, the company did not inform pilots about MCAS functionality until after the Lion Air accident (Committee on Transportation and Infrastructure, 2019, p. 6).

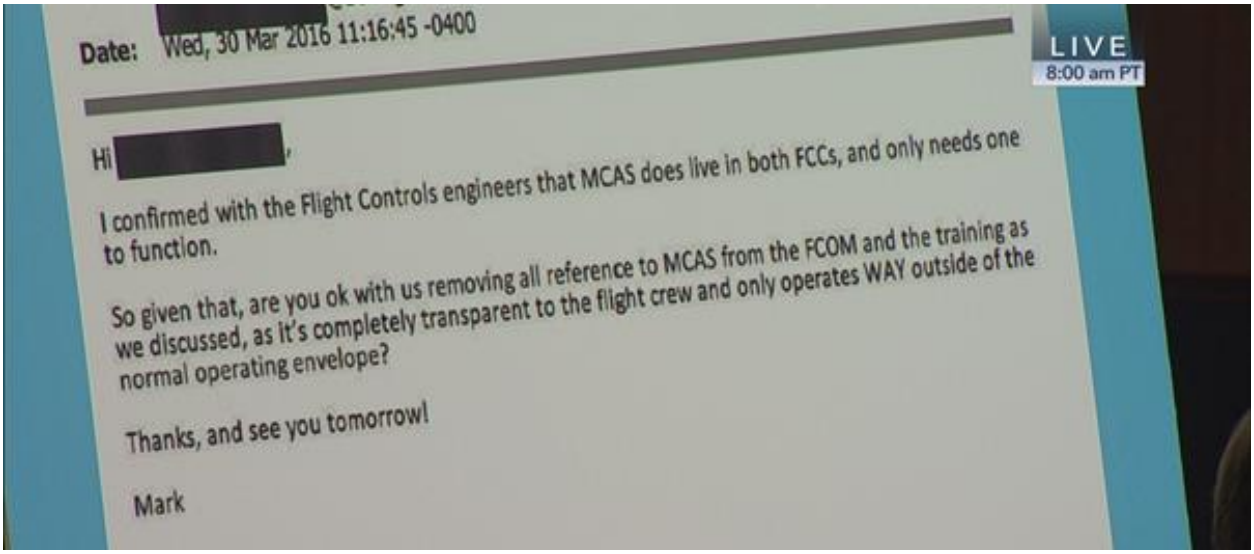

Figure 2. Email correspondence between system developers and the regulator demonstrating that seeing-through transparency was an intentional design choice for MCAS. Presented at the Senate committee meeting on aviation safety and the future of Boeing's 737 MAX.

It is possible, although not well-documented in the accident reports and U.S. Congressional hearings, that Boeing designers were concerned about the number of indications provided to pilots about automation systems they were not expected to have to interact with (The Air Current, 2018). As we noted in HWR-1250, a radical seeing-into approach to advanced automation will almost certainly result in information overload for operators.

That said, relying on seeing-through transparency had catastrophic consequences in this case. From a safety perspective, the failure of the design approach resulted in the deaths of 364 people, the loss of two airframes, and a worldwide safety shutdown of the aircraft. In economic terms, the grounding of the B737 MAX fleet will cost Boeing around $19 billion USD, including "...increased costs, loss of sales and revenue, loss of reputation, victims litigation, client compensation, decreased credit rating and lowered stock value" (Financial impact of the Boeing 737 MAX groundings, n.d.).

3.3 Implications for the Nuclear Industry

The Lion Air and Ethiopian Air B737 MAX accidents demonstrate that applying the inappropriate automation transparency principle may have substantial safety implications. Making the wrong design decision could also have colossal economic impacts if operational safety is compromised. Automation transparency is therefore not a hang-up of system designers or of mere academic interest. Rather, it is a real and present industrial automation design challenge.

The B737 MAX case sheds light on how regulatory regimes can influence design practices in unanticipated ways. The FAA (the responsible regulator) worked with Boeing to certify the new aircraft as a variant of an aircraft with an enviable safety record. By characterizing MCAS as an addition to the speed trim system, Boeing took explicit steps to avoid increased FAA certification. Their meeting minutes warned: "If we emphasize MCAS is a new function there may be a greater certification and training impact." (Committee on Transportation and Infrastructure, 2019).

Ultimately, the FAA condoned Boeing's adoption of the seeing-through approach to MCAS. Regulators in the nuclear industry should be alert to how the seeing-through approach to transparency might appeal to automation technology designers more focused on functionality than operator understanding of that functionality.

The B737 MAX case is a cautionary one; not a refutation of the seeing-through design approach. Under other operational circumstances, seeing-into transparency may overload operators and distract them from core tasks (as demonstrated in a HAMMLAB experiment from 2009; see Skraaning and Jamieson [2019]). In many situations, seeing-through transparency might be advisable. We observe this clearly in the automotive industry. Electronic Stability Control automatically and silently compensates for loss of traction during braking and steering. This feature saves thousands of lives around the world every year (Iombriller et al., 2019; Starnes, 2014). There would seem to be little technical basis for implementing such a system by adopting a seeing-into transparency approach. One may also speculate about whether the seeing-through transparency implementation for MCAS would have been a safe and reliable option if automation was activated based on consistent input from several AOA sensors, if maintenance practices for AOA sensors were less prone to failure, if the B737 MAX cockpit interface design did not overload pilots and obscure problems etc. (KNKT, 2019). Perhaps seeing-through transparency can be effectively realized if automation designers are able to consider the broader context of operation while at the drafting table.

The B737 MAX case raises a pressing research question for future nuclear plants: Are there generalizable problem characteristics that point to when the design of automation system should adopt the seeing-into or seeing-through transparency principles? And what implications does a choice of automation transparency principle have for the broader systems-engineering approach, including design for maintenance, system and software testing, training regimes, and regulatory review.

What we can deduce from the B737 MAX case study is a caution against the temptation to adopt one or the other principle for one of many attractive outcomes. These include realizing short-term economic gains; satisfying the hubris of automation designers; complying with corporate visions, technology trends, traditions, or industry culture; overemphasizing operator preferences; complying with anachronistic regulatory requirements; etc. The lesson learned from the B737 MAX accidents is that such critical design decisions should be knowledge-based and carefully validated in realistic test situations with humans in the loop.

4. TRANSPARENCY IN EXPLAINABLE ARTIFICIAL INTELLIGENCE

The B737 MAX automation transparency case study arises from a regulated, safety-critical domain with a considerable history of human-automation interaction research and practice. The MCAS automation is founded in familiar systems-engineering thinking and relies on system components (i.e., sensors, software, actuators) and environmental interactions that are familiar to nuclear processes.

We turn now to a distinctly different technology sector where notions of transparency are in vigorous use. Artificial Intelligence refers to a wide-ranging effort to develop machines that exhibit characteristics of human cognition; particularly learning and problem solving as expressions of centralized information processing. Explainable artificial intelligence (XAI) deals with the problem of allowing humans to understand and collaborate with such advanced forms of automation. We briefly introduced XAI in HWR-1250, and the literature has evolved significantly even in the past year (e.g., Arrieta et al., 2020). In

this report, we offer a broad, high-level account of the transparency notion arising from a dynamic XAI literature.

There are several compelling reasons for the nuclear industry to be aware of the transparency concepts arising from XAI. First, XAI is a discipline originating from computer science (in contrast to the engineering and psychology underpinnings of human factors and human-automation interaction research). This disciplinary pivot offers fresh thinking on the transparency of future intelligent systems. In particular, we are less knowledgeable about the nature of the technology, the computer programming, and the design processes employed in this discipline. Despite these limitations, it is still important to transcend that disciplinary boundary because the smartest and most complex forms of automation developed so far—at least from a programming perspective—are arising from computer science. These include deep learning algorithms that learn progressively from experience and organize knowledge in layers (artificial neural networks) similarly to humans (LeCun, Benigo, and Hinton, 2015). Thus, a second reason to be aware of XAI is that these are extreme technologies that may help us understand the future challenges and limits of automation transparency and project those for the nuclear industry. A third motivation for interest in XAI is that AI-based decision support and action implementation systems have been—and are being—adopted in other safety-critical industries. They are, for example, paving the way for driving automation, are embraced and explored in air traffic control (Eurocontrol, 2019), and are already established in healthcare and military command and control. Moreover, these adoptions have not necessarily been smooth.

The top ten industrial use cases listed in Figure 3 may also suggest how intelligent automation may benefit the future operation of nuclear plants.

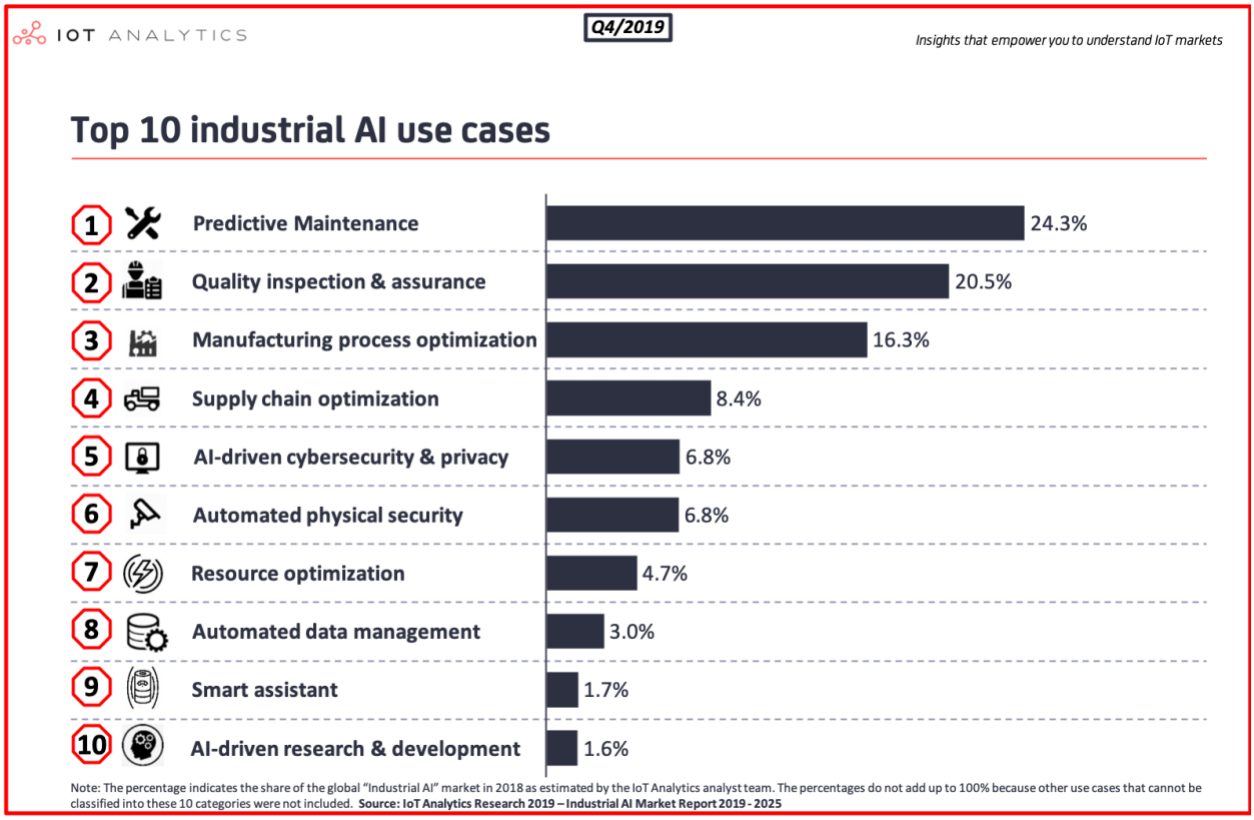

Figure 3. Top ten industrial AI use cases according to the Internet of Things (IoT) Analytics' 2020–2025 Market Report (the figure is from the IoT Analytic' web site, 2019).

4.1 Uses of Transparency in XAI

Our observation to date is that transparency is used inconsistently within the XAI community. Some authors (Mohseni et al., 2020) refer to transparency as disclosure of goals, activities, inner workings, and performance of intelligent systems (i.e. akin to seeing-into transparency as in human factors, human-automation interaction, and HRI). Other authors (e.g., Arrieta et al., 2020) use transparency more narrowly to characterize intelligent systems that are interpretable by design (so called glass box models). In contrast, opaque intelligent systems (black box models, e.g., deep learning algorithms) have to be "explained" post hoc. We concentrate on these latter interpretations of transparency and post hoc explanations as they add a new facet to our thinking about the disclosure decisions made by designers.

To make intelligent systems understandable to humans, AI-developers may, as suggested above, either 1) construct interpretable models that are transparent by design or 2) develop more complex opaque models that have to be explained post hoc, using dedicated techniques such as model simplification, assessment of "influence, relevance, and importance of model features for the predictability of output", and/or visualizations of the model (Arrieta et al., 2020, p. 19–20). Both the inherently transparent and post hoc explainable approaches disclose information about automation and thereby facilitate seeing-into transparency.

A key distinction in XAI, however, has been that both the interpretable model and post hoc explanation approaches are meant to disclose the inner workings of intelligent systems to AI-developers specifically, i.e. these methods were originally not intended for end users. That said, the goals, concepts and core thinking of XAI may very well be applicable to both developers and end users. It is also worth noting that there seems to be a recent change of attitude where the purpose of XAI is understood more in terms of end user than developer needs. For example, the objective of Defense Advanced Research Projects Agency (DARPA's) Explainable Artificial Intelligence program is to "create a suite of new or modified ML techniques that produce explainable models that, when combined with effective explanation techniques, enable end users to understand, appropriately trust, and effectively manage the emerging generation of AI systems" (Gunning and Aha, 2019, p. 45).

Unexplained intelligent systems can be argued to resemble seeing-through transparency, although such black box models are often incomprehensible due to technological immaturity or complexity. Thus, their opacity is not a feature, whereas seeing-through transparency may be an intentional design philosophy (e.g., in remote operations).

4.2 How Insights from XAI can Help the Nuclear Industry Address Future Human-Automation Interaction Challenges

We have noted how the XAI discipline characterizes intelligent systems as transparent by design or post hoc explainable. This distinction has inspired an extension of the automation transparency framework presented in HWR-1250 (see below) and pushed us towards a deeper understanding of the operational dilemmas associated with future forms of automation in nuclear plants.

We anticipate that such learning systems will appear in the nuclear power domain in the form of intelligent decision support agents, smarter alarm/diagnosis systems, handling of within-design basis events, operation/procedure automation in normal plant states, or even accident management/safety systems in future NPPs. Alternatively, and perhaps initially, advanced AI technologies are expected to be an integral part of microreactor development.

On the one hand, these anticipated technology incursions prompt an array of difficult questions. Is the nuclear industry ready to adopt the emergent complexities of future intelligent systems if automation becomes incomprehensible to system designers and plant operators? Will the industry be willing to implement stochastic algorithms that have to be justified and explained in a simplified manner depending

on the process output? Can we trust smarter and more autonomous automation that is also vague and semireliable?

On the other hand, what are the consequences for the future of our industry if the most intelligent automation is collectively rejected while other safety-critical sectors and competing power generating industries embrace the new AI technology? Is it a realistic alternative to rely on deterministic automation that is sufficiently intelligent, transparent by design, and thereby fully understandable to developers and operators? The downside of a conservative approach to the adoption of intelligent automation is that algorithmic complexity and intelligence tend to be positively correlated (i.e. smarter systems that may be necessary to remain competitive and relevant as an industry have so far been inscrutable by nature).

If the industry decided to open up to more opaque forms of automation, perhaps for operator support systems as an initial step, to what degree and in what sense should operators be expected to understand these new forms of intelligent automation? Based on the automation transparency framework presented in HWR-1250 and the discussion on XAI above, we believe that the depth of operator comprehension of automation in future plants could fall into three categories:

- *Full comprehension*: operators have a deep functional understanding of the inner workings of automation
- *Attainable comprehension*: operators are provided with the means to develop a simplified rationale in order to justify and build trust in the decisions and actions of otherwise opaque automation
- *Minimal comprehension:* seeing-through transparency where automation appears invisible to operators.

Whereas the full and minimal comprehension approaches have featured already in our automation transparency framework, attainable comprehension would require an extension. Given the likely opaque nature of smart automation in future plants, this might be a form of automation transparency that we should seek to develop. In the next section, we consider this from the perspective of industrial design approaches.

5. OVERVIEW OF AUTOMATION TRANSPARENCY DESIGN APPROACHES

Creating user-centred automation has challenged developers for decades, and this report highlights automation technology trends that deepen this challenge. As these technologies penetrate the nuclear industry, developers will seek out proven transparency design approaches to guide the development of interactive automation. We have discovered a breadth of design perspectives that overlap and intersect with the automation transparency principle. Figure 4 offers an initial taxonomy of automation transparency design approaches that adopt an end user orientation.

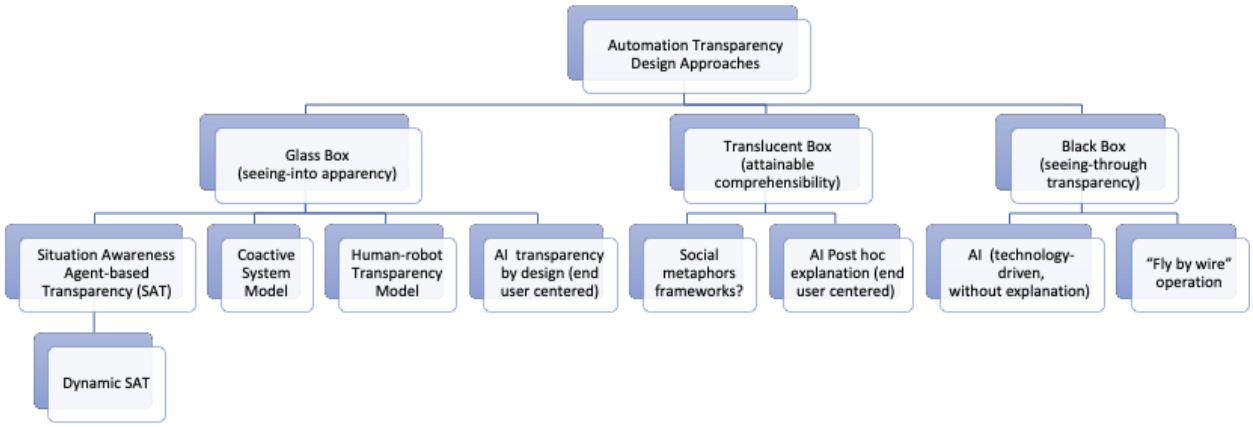

Figure 4. A taxonomy of (end user focused) automation transparency design approaches. Note the reintroduction of "apparency."

In describing Figure 4, we first emphasize that the taxonomy refers to end user automation design approaches. Additional perspectives on transparency emerge from developer-oriented methods in XAI that fall outside of the project scope (e.g., Mohseni et al., 2018).

The first level of the hierarchy distinguishes between three approaches to automation design and aligns each approach with a "box" metaphor. Much of the transparency and XAI literature adopts the black box versus glass box metaphor and we have found that metaphor to be reasonably robust. The notion of a glass box aligns with the seeing-into concept where automation (i.e., the contents of the box) is visible to operators[7]. The black box concept is analogous to a design approach where the inner workings of the automation are hidden from the operator (with the exception of inputs and outputs)[8]. Inspired by our reading of the XAI literature, we have extended the metaphor to include a translucent box; one that can be peered into to reveal useful but limited insight into its contents.

At the second level of the hierarchy are the design models, frameworks and methods that exemplify the higher-level frameworks. Under the glass box metaphor, we find the SAT model and its expansion as the Dynamic SAT model (Chen et al., 2018), the Coactive System Model (Johnson et al., 2014), and the Human-robot Transparency Model (Lyons, 2013). Also included under the glass box metaphor are AI methods that seek inherently interpretable models. This list could be expanded with additional models that have yet to gain traction in the literature.

A similar expansion of the Translucent Box can be developed. Although we have not identified any explicit Translucency (again, a label that we are introducing here) models, frameworks, or methods, it is plausible that transparent counterparts could be adapted to the more modest design objectives. Another possible direction for the development of translucent automation would be through application of social metaphors, which we explored in HWR-1128. Whereas we have questioned the future viability of the venerable human supervisory control metaphor, substantial insight into apparency requirements might be gained from other contexts in which humans interact closely with intelligent agents. These can include both human-human (e.g., the Butler metaphor) and human-animal (e.g., the horse and rider, the hunting dog). Social metaphors encapsulate many expectations about the knowledge, skills, and awareness of—

7 Visibility does not necessarily imply controllability. A glass box might make automation apparent without affording accessibility in terms of control.

8 Although the glass box and black box metaphors align well with the seeing-into and seeing-through uses of automation transparency, we employ both sets of metaphors here to motivate the "translucent box" metaphor; a new insight arising from our encounter with the XAI community.

and transactions between—agents. However, these metaphors are both approximate and incomplete; characteristics that might approximate what we mean by translucent automation.

Under the black box metaphor, we have denoted two approaches. The first includes methods like deep learning, which are inherently intractable to human understanding and thus, by design, foreign to the notion of explanation. We distinguish these from control methods that, while wholly modellable by engineers, are sufficiently robust to be entirely hidden from the end user. These methods arise from the teleoperation tradition and are elegantly expressed in "fly-by-wire" technologies arising from the aviation domain.

5.1 Implications

We anticipate that these myriad perspectives on end user (and developer) oriented automation design approaches could inform the development of a practical roadmap for the implementation of automation transparency in future nuclear facilities. Such a roadmap could aid developers in selecting and practicing an automation transparency approach, and it could aid regulators in reviewing the suitability of a chosen approach and verifying that it was consistently applied. However, it is worth noting that there is disagreement within the XAI community, and between it and the human factors community, about the organization of these approaches. Varieties in the use of language, design objectives, audience (i.e., to whom should the automation be transparent), content of communication, etc., make these approaches difficult to categorize and compare.

If there is to be a standardized taxonomy for the implementation of automation transparency, it is yet to be established—and obtaining consensus on such a framework across disciplines would likely be a slow and contentious process. It is not our goal to resolve these perspectives. Rather, we call attention to the plausible situation where designers working within one disciplinary context would not have broad awareness of alternative design approaches and the use of terms in other disciplines. This could readily lead to miscommunication about design intent and thus hamper verification and validation efforts.

6. CONCLUSION: INFORMING AND ADVANCING LWRS PROGRAM GOALS THROUGH AUTOMATION TRANSPARENCY

The objective of DOE's LWRS Program is to support the long-term sustainability of U.S. commercial NPPs. LWRS Program researchers conduct R&D to modernize technologies and improve processes, thereby providing the technical bases that help reduce the uncertainty and risk of full plant modernization. This report addresses the potential uncertainties and risks associated with the introduction of automation technology to the plant and operating environment as a means to fulfill the purpose of the LWRS Program objectives. We specifically concentrate on the ability of control room crews to understand the inner workings of automation or to appropriately rely on capable automation operating in the background.

The preceding investigation supports several conclusions:

1. The Boeing 737 MAX case study establishes that *automation transparency is an industrial safety problem*. This and other operating experiences with advanced automation in safety-critical industries offer a significant learning opportunity for the nuclear power industry.
2. To date, the most concerted efforts to develop a seeing-into transparency design framework have suffered from inconsistency in design expression and empirical validation. Other efforts yield alternative frameworks that are largely untested. The implication of these observations and findings is that the technical basis for seeing-into automation transparency is insufficient to inform design.
3. Realistic testing of transparency design solutions is necessary to make informed design and technology acceptance decisions. There is no known analytical alternative to such empirical testing at present.

4. The seeing-into automation transparency research base is almost exclusively situated in military command and control. While significant insight can be extracted from this research, the NPP sector should consume these results with caution.
5. XAI raises new challenges and suggests compelling alternative perspectives on automation transparency. However, this literature offers no clear path forward with respect to the preceding challenges—and should be expected to introduce new challenges as new forms of automation emerge.
6. To the extent that technology developers might be aware of the contrasting seeing-into and seeing-through design principles, there is no existing guidance regarding the selection of these principles—and little guidance regarding their application (see points 1 and 2 in this list).

Based on our research to date, we anticipate several challenges for the U.S. nuclear industry related to automation transparency.

1. Is the seeing-through design principle a viable alternative for developers of NPP control room automation? Is it plausible that the regulator would deem acceptable a technology that eludes the understanding of operating crews? Would blind trust and reliance be considered sufficient validating evidence?
2. What are the uncertainties and risks associated with adopting a translucent automation design principle where operators are informed on a need to know basis to facilitate trust and (where necessary) reliance without a full understanding of the automation's inner workings?
3. What would a translucent automation design framework look like? Perhaps a measured application of an existing seeing-through transparency framework (although see point 1 in this list)?
4. What expectations could be established for verification and validation of a translucent technology?
    a) What criteria should designers use in selecting between the seeing-into, translucent, or seeing-through principles?
    b) How could a technology developer or regulator verify that a given transparency framework has been chosen and employed?
5. How could a developer or regulator validate that a specific technology change to a plant or operating environment meets the expectations for crew performance in terms of automation transparency?

It is evident that control room automation will comprise a significant portion of the I&C technologies that ensure the long-term sustainability of the U.S. light-water reactor fleet. Decades of operating experience and human factors research suggest that the design decisions about the roles of operating crews in interacting with this automation will determine the safe and effective operation of more highly automated facilities. Moreover, emerging technologies, such as intelligent agents, will complicate these decisions and raise the stakes for verification and validation efforts. Automation transparency principles offer a useful approach to thinking about those decisions and anticipating their impacts. However, there are many gaps in the automation transparency operating experience and academic literature that point to uncertainties and risks that can be resolved and mitigated through further R&D efforts.

7. REFERENCES

Aircraft Accident Investigation Bureau, AAIB. 2019. *Preliminary Aircraft Accident Investigation Report, Boeing 737-8 (MAX), Ethiopian Airlines flight 302*.

Arrieta, A. B., N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, and R. Chatila. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI." Information Fusion 58: 82–115.

BBC. 2019. “Boeing Safety System Not at Fault, Says Chief Executive.” BBC, April 30, 2019 (Accessed July 8, 2020). https://www.bbc.com/news/business-47980959.

Bhaskara, A., M. Skinner, and S. Loft. 2020. “Agent Transparency: A Review of Current Theory and Evidence.” IEEE Transactions on Human-Machine Systems 50(3): 215–224.

Chen, J. Y., S. G. Lakhmani, K. Stowers, A. R. Selkowitz, J. L. Wright, and M. Barnes. 2018. “Situation Awareness-Based Agent Transparency and Human-Autonomy Teaming Effectiveness.” Theoretical Issues in Ergonomics Science 19(3): 259–282.

The House Committee on Transportation and Infrastructure. 2020. “The Boeing 737 MAX Aircraft: Costs, Consequences, and Lessons from Its Design, Development, and Certification.” March 2020 (Accessed 2020-06-11). https://www.rapoportlaw.com/TI-Preliminary-Investigative-Findings-Boeing-737-MAX-March-2020.pdf.

Endsley, M. R. 2019. “Human Factors & Aviation Safety. Testimony to the United States House of Representatives Hearing on Boeing 737-Max8 Crashes.” Human Factors and Ergonomics Society, December 2019 (Accessed 2020-06-07). https://higherlogicdownload.s3.amazonaws.com/HFES/42fffbb4-31e1-4e52-bda6-1393762cbfcd/UploadedImages/Human_Factors_and_the_Boeing_737-Max8-FINAL.pdf.

Eurocontrol, 2019. Why artificial intelligence is highly relevant to air traffic control. Accessed 2020-08-11. https://www.eurocontrol.int/article/why-artificial-intelligence-highly-relevant-air-traffic-control

Gunning, D., and D. W. Aha. 2019. “DARPA’s Explainable Artificial Intelligence Program.” AI Magazine 40(2): 44–58.

Guznov, S., J. Lyons, M. Pfahler, A. Heironimus, M. Woolley, J. Friedman, and A. Neimeier. 2020. “Robot Transparency and Team Orientation Effects on Human–Robot Teaming.” International Journal of Human–Computer Interaction 36(7): 650–660.

Iombriller, S. F., W. B. Prado, and M. A. Silva. 2019. “Comparative Analysis between American and European Requirements for Electronic Stability Control (ESC) Focusing on Commercial Vehicles.” No. 2019-01-2141, SAE Technical Paper.

IOT Analytics. 2019. “The Top 10 Industrial AI Use Cases.” Accessed July 8, 2020. https://iot-analytics.com/the-top-10-industrial-ai-use-cases/.

Jamieson, G. A., and G. Skraaning. 2020. “The Absence of Degree of Automation Trade-Offs in Complex Work Settings.” Human Factors 62(4): 516–529.

Johnson, M., J. M. Bradshaw, P. J. Feltovich, C. M. Jonker, M. B. Van Riemsdijk, and M. Sierhuis. 2014. “Coactive Design: Designing Support for Interdependence in Joint Activity.” Journal of Human-Robot Interaction 3(1): 43–69.

Komite Nasional Keselamatan Transportasi, KNKT. 2019. *Final Aircraft Accident Investigation Report, Boeing 737-8 (MAX), Lion Air flight 610*.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. “Deep Learning.” Nature 521(7553): 436–444.

Skraaning Jr., G., and G. A. Jamieson. 2019. “Human Performance Benefits of the Automation Transparency Design Principle: Validation and Variation.” Human Factors. https://doi.org/10.1177/0018720819887252.

Lyons, J. B. 2013. “Being Transparent about Transparency: A Model for Human-Robot Interaction.” Paper presented at the 2013 AAAI Spring Symposium Series, Stanford, CA, March 2013.

Mercado, J. E., M. A. Rupp, J. Y. C. Chen, M. J. Barnes, D. Barber, and K. Procci. 2016. "Intelligent Agent Transparency in Human-Agent Teaming for Multi-UxV Management." Human Factors 58(3): 40–415. http://doi.org/10.1177/0018720815621206.

Mohseni, S., N. Zarei, and E. D. Ragan. 2018. "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems." arXiv, arXiv–1811.

Oxstrand, J. H., and K. Le Blanc. 2012. "Computer-Based Procedures for Field Workers in Nuclear Power Plants: Development of a Model of Procedure Usage and Identification of Requirements." INL/EXT-12-25671, Idaho National Laboratory.

Selkowitz, A. R., S. G. Lakhmani, C. N. Larios, and J. Y. Chen. 2016. "Agent Transparency and the Autonomous Squad Member." In the Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 60, no. 1, 1319–1323. Los Angeles, CA: SAGE Publications.

Sheridan, T. B., and W. L. Verplank. 1978. "Human and Computer Control of Undersea Teleoperators." Massachusetts Inst of Tech., Cambridge, Man-Machine Systems Lab.

Skraaning, G., and G. A. Jamieson. 2019. "Human Performance Benefits of the Automation Transparency Design Principle: Validation and Variation." Human Factors, https://doi.org/10.1177/0018720819887252.

Skraaning, G., G. A. Jamieson, F. Rajabiyazdi, and N. Mirjalali. 2019. "New Angles on Automation Transparency." HWR-1250, OECD Halden Reactor Project, Halden, Norway.

Starnes, M. 2014. "Estimating Lives Saved by Electronic Stability Control, 2008–2012." Report No. DOT HS 812 042, National Highway Traffic Safety Administration.

Stowers, K., N. Kasdaglis, O. Newton, S. Lakhmani, R. Wohleber, and J. Chen. 2016. "Intelligent Agent Transparency: The Design and Evaluation of an Interface to Facilitate Human and Intelligent Agent Collaboration." In the Proceedings of the Human Factors and Ergonomics Society, vol. 60, issue 1, 1706–1710. http://doi.org/10.1177/1541931213601392.

The Air Current. 2019. "What Is the Boeing 737 MAX Maneuvering Characteristics Augmentation System?" Accessed 2020-06-11. https://theaircurrent.com/aviation-safety/what-is-the-boeing-737-max-maneuvering-characteristics-augmentation-system-mcas-jt610/.

Thomas, K., R. Boring, R. Lew, T. Ulrich, and R. Vilim. 2013. "A Computerized Operator Support System Prototype." INL/EXT-13-29651, Idaho National Laboratory.

U.S. Congressional hearing on the B737 MAX accidents. 2019. Accessed 2020-29-06. https://www.youtube.com/watch?time_continue=4&v=IoK0D3FDTnE&feature=emb_logo5:21:20.

Wikipedia. n.d. "Financial Impact of the Boeing 737 MAX Groundings." Accessed June 30, 2020. https://en.wikipedia.org/wiki/Financial_impact_of_the_Boeing_737_MAX_groundings.

Wright, J. L., J. Y. Chen, M. J. Barnes, and P. A. Hancock. 2017. "Agent Reasoning Transparency: The Influence of Information Level on Automation Induced Complacency." No. ARL-TR-8044, U.S. Army Research Laboratory Aberdeen Proving Ground, United States.

Wright, J. L., J. Y. Chen, M. J. Barnes, and P. A. Hancock. 2016. "Agent Reasoning Transparency's Effect on Operator Workload." In the proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 60, no. 1, 249–253. Los Angeles, CA: SAGE Publications.