Derek Rayside with Emily Huang & Ayush Kapur

An Engineering View of Regulating AI in Canada's Financial Institutions

University of Waterloo AI Institute
Cheriton School of Computer Science Technical Report #CS-2020-02
https://cs.uwaterloo.ca/research/technical-reports/2020-technical-reports
February 28, 2020

Executive Summary

Canada's financial institutions increasingly see themselves as technology companies, and are eager to apply artificial intelligence (AI). Technology companies are sometimes known to "move fast and break things," which is not consistent with the long tradition of prudent management in Canada's financial institutions.

Autonomous vehicles represent an example of professionally responsible engineering of AI technologies, generally believed to be in the interest of improving public safety. Perhaps it would be interesting for financial institutions to consider what is happening here.

There is a high level of collaboration in autonomous vehicles between corporations, professional bodies (primarily SAE, the Society of Automotive Engineers), and government regulators (including, NHTSA, the US National Highway Transportation Safety Administration). NHTSA has adopted standards proposed by SAE (such as J3016), and has worked through multiple rounds of consultation on developing their *Proposed Voluntary Guidance for Autonomous Vehicles.*

www.sae.org
www.nhtsa.gov

https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety

The core of this report is an adaptation of NHTSA's guidance to the context of Canada's financial institutions. The basic idea of the adaptation is to take NHTSA's text and change 'car company' to 'financial institution', 'advanced driving system' to 'advanced financial system', and so on. This gives a view of what guidance for the design and deployment of AI technologies in Canada's financial institutions might look like — at least it illustrates a professionally responsible engineering approach to AI.

Three important engineering concepts in this report that might be interesting for financial institutions include *Operational Design Domain* (ODD), *fallback to minimal risk*, and *human-machine interaction*. Technical aspects of *bias*, *explainability*, and *data* are also discussed.

This report was stimulated by a workshop hosted a the University of Waterloo's AI Institute in autumn 2019, at which representatives from OSFI (Office of the Superintendent of Financial Institutions) asked Waterloo researchers for their thoughts on AI. In this report, NHTSA's text is expanded in several places based on ideas discussed by these researchers.

www.waterloo.ai
www.osfi-bsif.gc.ca

OSFI did not commission this report. This report does not describe anything that OSFI has done or will do. The opinions contained herein are those of the authors and not of OSFI.

Contents

1 *Preface*

This report adapts guidance for advanced automation technologies (*i.e.,* AI) from the context of *autonomous vehicles* to the context of Canada's *financial system*. The main body of the report is an adaptation of the American *National Highway Traffic Safety Administration*'s (NHTSA) proposed voluntary guidance for vehicle manufacturers regarding use of advanced automation technologies. Within the automotive industry, NHTSA's approach is relatively well regarded as finding a good balance between encouraging innovation and their regulatory mandate for ensuring public safety on the nation's highways. NHTSA believes, with empirical evidence, that advanced automation technologies can improve road safety, and so encouraging the appropriate development and deployment of such technologies is in the public interest and within their regulatory mandate.

NHTSA has gone through three rounds of public consultation and revision of their guidelines over the last five years. The version 2 guidance completely replaced the version 1 guidance. The latest, version 3 guidance, is a refinement of the version 2 guidance. This is one view of what successful guidance around AI might look like in an area that is crucial for society to function effectively, where advanced automation technologies hold great promise but also potential for misuse and negative consequences.

1.1 NHTSA's *Motivation and Process*

In NHTSA's own words:

https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety

> In September 2016, NHTSA and the U.S. Department of Transportation issued the Federal Automated Vehicles Policy which set forth a proactive approach to providing safety assurance and facilitating innovation. Building on that policy and incorporating feedback received through public comments, stakeholder meetings, and Congressional hearings, in September 2017, the agency issued, Automated Driving Systems: A Vision for Safety 2.0. The updated guidance, 2.0, offers a flexible, nonregulatory approach to automated vehicle technology safety, by supporting the automotive industry and other key stakeholders as they consider and design best practices for the safe testing and deployment of ADS levels 3 through 5. It also provides technical assistance to states and best practices for policymakers regarding ADS.
>
> In October 2018, U.S. DOT released Preparing for the Future of Transportation: Automated Vehicles 3.0, which builds upon — but does not replace — the voluntary guidance provided in 2.0. AV 3.0 expands the scope to all surface on-road transportation systems, and was developed through input from a diverse set of stakeholder engagements throughout the nation. AV 3.0 is structured around three key areas:
>
> - Advancing multi-modal safety,
> - Reducing policy uncertainty, and
> - Outlining a process for working with U.S. DOT.
>
> The U.S. DOT sees AV 3.0 as the beginning of a national discussion about the future of our on-road surface transportation system. As automated technologies advance, so will the department's guidance. The guidance is intended to be flexible and to evolve as technology does, but with safety always as the top priority.

https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/documents/13069a-ads2.0_090617_v9a_tag.pdf

https://www.transportation.gov/sites/dot.gov/files/docs/policy-initiatives/automated-vehicles/320711/preparing-future-transportation-automated-vehicle-30.pdf

This report is based largely on the language of NHTSA's AV2.0, which in our judgement translates more readily to the context of Canada's financial institutions.

1.2 Considering Financial Institutions as Technology Companies

Canada's Financial Institutions increasingly consider themselves technology companies. This is certainly the image they portray when they come to campus to hire students educated in STEM-related disciplines. As technology companies, they are in the business of developing and deploying advanced automation systems, including those based on AI techniques such as *neural networks*.

FI: Financial Institution

STEM: Science, Technology, Engineering, Mathematics

AI: Artificial Intelligence

1.2.1 Engineering Professional Practice

It is our professional engineering opinion that this report is relevant and useful for Canada's Financial Institutions insofar as they are technology companies. Engineering proceeds, in part, by learning from experience. This time-tested tradition is part of how engineers protect the public interest.

The essence of NHTSA's guidelines are good engineering professional practice, as would be taught in an Engineering 101 course. To us, it appears obvious that this wisdom is also applicable with regards to the development of advanced automation technology in the financial domain — just as it is for autonomous vehicles and other areas of engineering. In our adaptation of NHTSA's guidelines, we have left their language about professional engineering practice in tact. Where NHTSA's guidelines have language specific to autonomous vehicles, we have adapted it to the context of Canada's financial system as best we can.

1.2.2 Move Fast and Break Things

It is our personal opinion, as individual citizens and consumers who use services provided by Canada's Financial Institutions, that our confidence and trust in the financial system is reassured by the thought that Canada's Financial Institutions would engage in reasonable and responsible professional engineering practices when working with advanced automation technologies. Our consumer confidence and trust is not reassured by the idea of technology companies who "move fast and break things." As consumers and citizens, our understanding is that Canada's Financial Institutions survived the 2008 crash reasonably well because of prudent professional management. We hope that this tradition of prudence and stability continues.

"Move fast and break things" is a phrase common with technology startups. It entered popular culture via Facebook CEO Mark Zuckerberg's use of it in Facebook's early days. https://hbr.org/2019/01/the-era-of-move-fast-and-break-things-is-over

1.3 OSFI *Workshop at UWaterloo's* AI *Institute*

In fall 2019, the University of Waterloo's AI Institute hosted a daylong workshop with academics and representatives from OSFI for a broad discussion of AI from a variety of perspectives. Academic researchers present at the workshop included, in alphabetical order:

http://waterloo.ai

OSFI: Office of the Superintendent of Financial Institutions

- Joel Blit, PhD, *Economics*
- Paul Fieguth, PhD, P.Eng., *Systems Design Engineering*
- Plinio Morita, PhD, P.Eng., *School of Public Health and Health Systems*
- Derek Rayside, PhD, P.Eng., *Electrical & Computer Engineering*
- Anindya Sen, PhD, *Economics*
- Alex Wong, PhD, P.Eng., *Systems Design Engineering*

This report attempts to integrate some of the ideas of these researchers within the structure of NHTSA's guidelines.

This report should not be construed as coming from OSFI. OSFI did not request this report and did not write this report. This is a report by academics and engineers who have limited knowledge and experience with finance and with OSFI, but who are experts in some areas of artificial intelligence and public policy, and who think the question of how to regulate the use of advanced automation technologies in Canada's financial system is interesting.

1.3.1 *About the Authors*

This report was written by Prof Derek Rayside with students Emily Huang and Ayush Kapur. It includes some ideas from other participants at the UW+OSFI Workshop, which are attributed — usually in the margin notes, but sometimes in the main text.

Derek is the Director of Software Engineering at the University of Waterloo, where he is also Faculty Advisor to the Watonomous autonomous vehicle team and a member of Waterloo's AI Institute. He is a licensed professional engineer in Ontario, an Associate Professor of Electrical & Computer Engineering at Waterloo, and is also cross-appointed to the Cheriton School of Computer Science.

Watonomous.ca

Waterloo.ai

Derek taught Emily and Ayush in Software Engineering 101. Emily and Ayush both enjoy studying software engineering, philosophy, and political science.

The Software Engineering program at the University of Waterloo is a collaboration between the Cheriton School of Computer Science, in the Faculty of Mathematics, and the Department of Electrical & Computer Engineering, in the Faculty of Engineering.

The University of Waterloo Artificial Intelligence Institute represents approximately 100 researchers across all six faculties at the University of Waterloo: Arts, Applied Health Science, Engineering, Environment, Mathematics, and Science.

Waterloo.ai

1.4 *Methodology of this Report*

The body of this report is a adaptation of NHTSA's proposed voluntary guidance to vehicle manufacturers on advanced automation systems in the context of autonomous vehicles. This report draws upon both NHTSA's *Automated Driving Systems 2.0: A Vision for Safety* and *Preparing for the Future of Transportation: Automated Vehicles 3.0* to incorporate the former's thorough voluntary guidance with the latter's updated clarifications and amendments. The methodology of this report is to copy NHTSA's language and modify where appropriate. The overall structure of NHTSA's guidance is preserved.

Much of NHTSA's language describes good engineering professional practice, as would be taught in an Engineering 101 course. This language is usually preserved as is.

Some of NHTSA's language is specific to the context of autonomous vehicles. Some of that language has been adapted, and some has been removed. There are also a few places where we have added ideas from the UW+OSFI Workshop. In some places, we have provided commentary in the margins about modifications that have been made to NHTSA's original text.

In most sentences, NHTSA's language has been preserved modulo some simple terminology substitutions, as described in Table 1.1. In a few cases, more editing was required to translate NHTSA's intentions from the automotive to the financial contexts, but efforts were made to preserve sentence structure and ordering.

NHTSA	*Substitution*	*Note*
Entity	FI	Financial Institution(s)
NHTSA	OSFI	the regulatory body
ADS	AFS	Automated Driving/Financial System(s)
state	province	political nomenclature
test-track	historical data	a testing context
on-road	human-monitored	testing the AFS in a live market
driver	operator	individual involved in decision making
occupant	consumer	individual outside the entity/FI

Table 1.1: Terminology substitutions to translate NHTSA's text from the automotive domain to the financial domain.

PLURAL FORMS OF ACRONYMS are for the reader to interpret as appropriate. Adding an 's' on to the end of an acronym to indicate plurality does not necessarily enhance readability. For example, sometimes 'FI' should be read as *Financial Institution*, and sometimes it should be read as *Financial Institutions*, depending on context.

Possessive forms of acronyms are indicated with an apostrophe and a trailing letter 's', as ordinarily expected.

WE HOPE that this report will stimulate new thoughts in the financial ecosystem. NHTSA's guidance does a good job of describing where the rubber meets the road, so to speak, in the design and deployment of advanced automation systems, and maybe that will help ground discussions in the financial world as the analogies are explored.

1.4.1 *Safety, Hazard, and Crash Metaphors*

NHTSA uses the terms *safety*, *hazard*, and *crash* frequently. Their mandate is highway safety. It is relatively clear what these terms mean in NHTSA's context, even though NHTSA does not define them extensively. Some aspects of transportation safety include:

- *Vehicle occupants:* physical safety of vehicle occupants, obviously.
- *Pedestrians:* physical safety of pedestrians and other vulnerable road users. The focus at this level is that all road users should be safe. Philosophers sometimes like to consider the ethical questions around prioritizing the safety of some classes of road users over others — this is sometimes referred to as *the trolley problem*.

https://en.wikipedia.org/wiki/Trolley_problem

- *Infrastructure:* AI shouldn't cause undue damage to the road itself.
- *Emergent Effects:* individual agents making locally optimal choices shouldn't lead the transportation system as a whole to a sub-optimal state. One kind of example of this kind of problem could be how a collection of local choices might concentrate overall system risk.
- *Public Trust:* Maintaining the public trust is an important aspect of advanced automation technologies, whether it be in the context of autonomous vehicles or in the context of the financial system.

What exactly *safety*, *hazard*, and *crash* mean in the context of Canada's financial system is left as an exercise to the reader. We think this is an important question that others are better qualified to answer than we are. We think it is thought provoking to consider these questions, and to consider how AI might improve 'safety' in the financial context. Some potential aspects of financial safety that come to our minds as academics include:

- Consumers should not be financially harmed by AFS.
- FI should not be exposed to exaggerated risk by AFS.
- AFS should preserve or enhance shareholder value.
- The greater financial ecosystem will hopefully become more robust and resilient to ordinary economic fluctuations through the deployment of AFS.
- The public's trust in the financial ecosystem will be maintained, and perhaps even enhanced, through the deployment of AFS.

1.5 *Comparison to other sources of guidance on* AI

There are many sources of guidance on development and deployment of advanced automation technologies (*i.e.*, AI). For example:

- *The Government of Canada Directive on Automated Decision-Making* https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592
- *The Government of Canada Algorithmic Impact Assessment Tool* https://www.canada.ca/en/government/system/digital-government/modern-emerging-technologies/responsible-use-ai/algorithmic-impact-assessment.html
- *The Canada Council on* AI https://www.canada.ca/en/government/system/digital-government/modern-emerging-technologies/responsible-use-ai.html
- *The Montreal Declaration for Responsible Development of Artificial Intelligence* https://www.montrealdeclaration-responsibleai.com/
- *The Toronto Declaration: Protecting the rights to equality and non-discrimination in machine learning systems* https://www.accessnow.org/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discriminati
- *Carlton University, Norman Paterson School of International Affairs Recommendations* https://www.cifar.ca/docs/default-source/student-symposium-ai-human-rights/aistudentsymposium_almeidaetal_presentation.pdf?sfvrsn=afaff48f_4
- *The Beijing AI Principles* https://www.baai.ac.cn/blog/beijing-ai-principles

Most of this other important guidance is of an ethical or legal nature, and was not written by an engineering regulatory body (with some exceptions). We expect the contents of this report to be consistent with and complementary to this other guidance.

Where we expect this report to have important and novel contributions is its holistic engineering perspective: this report speaks about how technology development and deployment should be managed in a professional and organizational context; moreover, this report is derived from guidance written by an engineering regulatory body (NHTSA), so it gives an example of what language from such a body might sound like. Vehicle manufacturers, like financial institutions, are large organizations with many people working together — sometimes closely and sometimes loosely — towards a set of common goals. This report, following NHTSA's lead, speaks in part to the professional practices these people should employ in the development and deployment of advanced automation systems, which includes ethical and legal concerns, but also other professional matters.

There are ways in which AI is new and unlike any previous technologies. Yet there are also ways in which it is similar to existing technologies, and NHTSA's example shows that some existing engineering professional practices remain relevant in the context of AI.

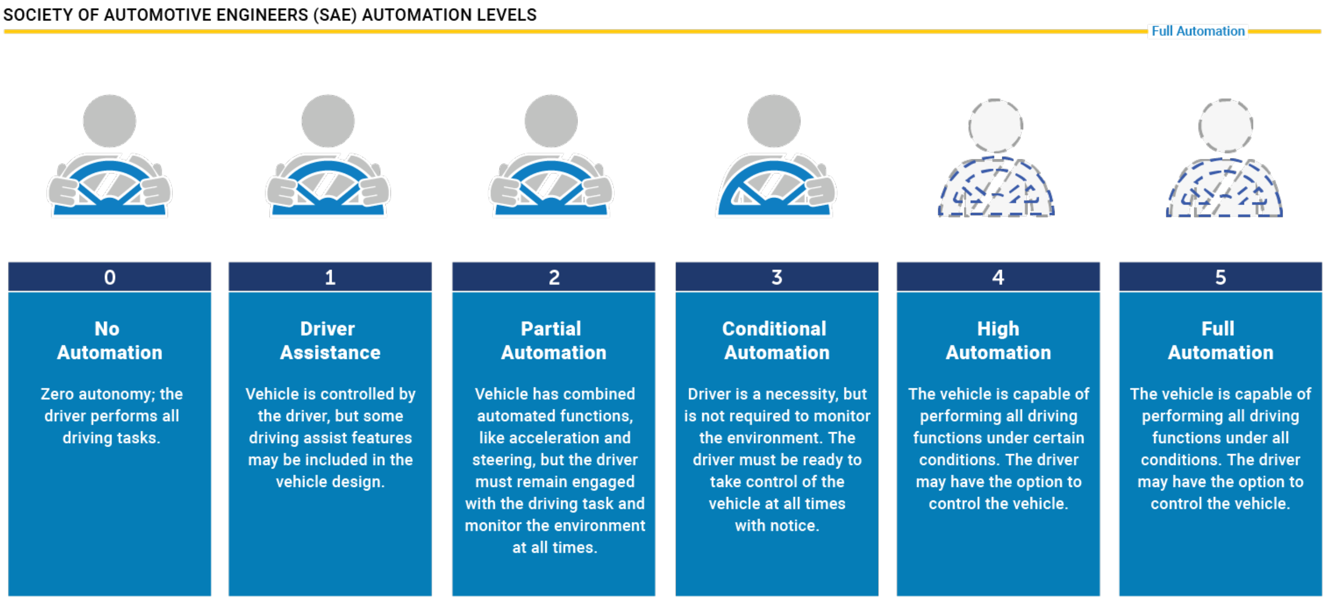
2

SAE *Levels of Automation*

SAE is the *Society of Automotive Engineers*, which is a professional body and a standards organization that represents engineers in three related industry segments: automotive, aerospace, and off-road (*e.g.*, farming and mining). SAE standard J3016, adopted by NHTSA, defines six different levels of automation, from *no automation* (level 0) to *full automation* (level 5), as described in Figures 2.1 and 2.2.

As Dr Gill Pratt, Chairman of Toyota Research Institute, has said over multiple years in his remarks at the *Consumer Electronics Show*, nobody really knows how to build level 5 in automotive: that is a goal being pursued, but not a reality of today. Today, vehicles for sale in North America are all at level 2 *partial automation*. Research technology exists for level 3 *conditional automation*, and arguably level 4 *high automation*, but that technology is not available for sale to consumers in North America.

https://global.toyota/en/newsroom/corporate/26085202.html



0 **No Automation** Zero autonomy; the driver performs all driving tasks.

1 **Driver Assistance** Vehicle is controlled by the driver, but some driving assist features may be included in the vehicle design.

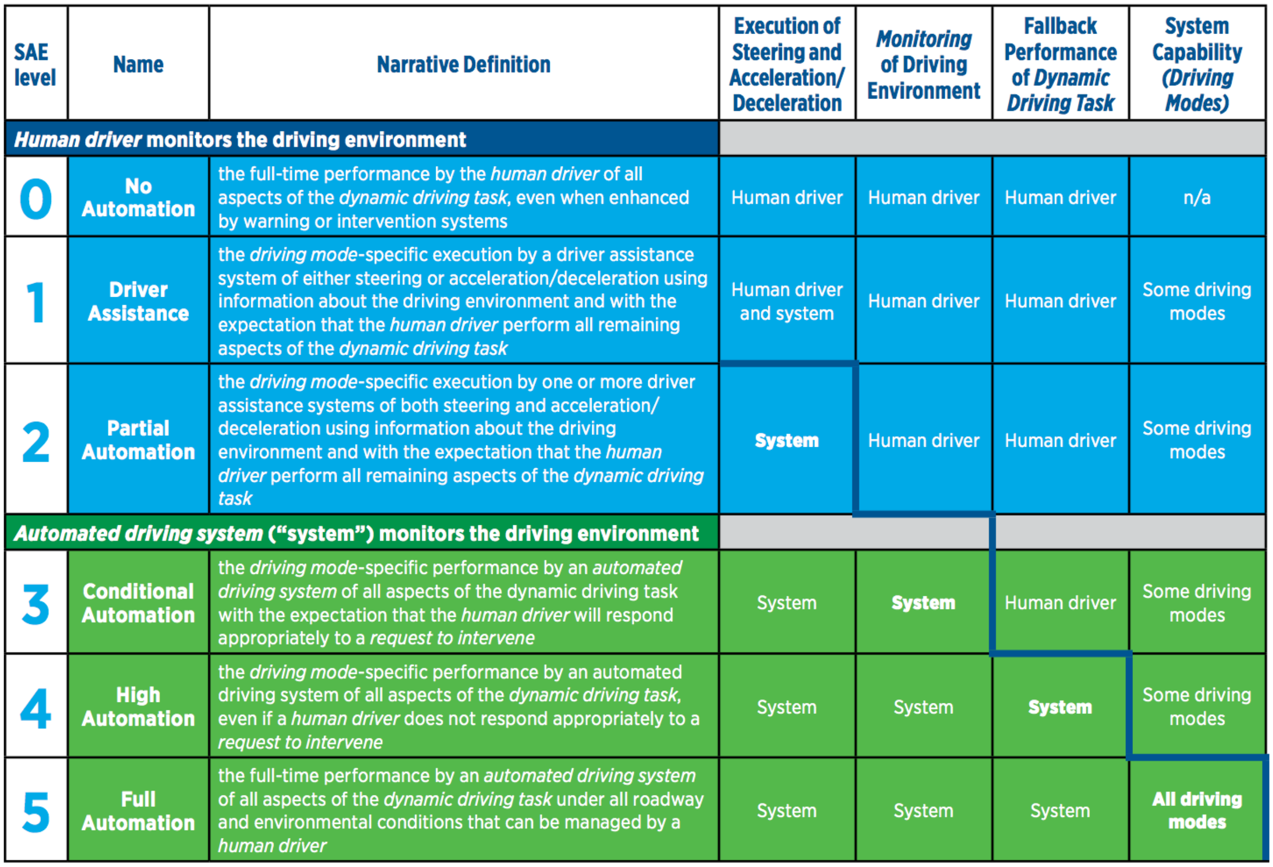
2 **Partial Automation** Vehicle has combined automated functions, like acceleration and steering, but the driver must remain engaged with the driving task and monitor the environment at all times.

3 **Conditional Automation** Driver is a necessity, but is not required to monitor the environment. The driver must be ready to take control of the vehicle at all times with notice.

4 **High Automation** The vehicle is capable of performing all driving functions under certain conditions. The driver may have the option to control the vehicle.

5 **Full Automation** The vehicle is capable of performing all driving functions under all conditions. The driver may have the option to control the vehicle.

Figure 2.1: SAE infographic on levels of automation, from NHTSA website https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety



SAE level	Name	Narrative Definition	Execution of Steering and Acceleration/ Deceleration	*Monitoring* of Driving Environment	Fallback Performance of *Dynamic Driving Task*	System Capability *(Driving Modes)*
Human driver **monitors the driving environment**						
0	**No Automation**	the full-time performance by the *human driver* of all aspects of the *dynamic driving task*, even when enhanced by warning or intervention systems	Human driver	Human driver	Human driver	n/a
1	**Driver Assistance**	the *driving mode*-specific execution by a driver assistance system of either steering or acceleration/deceleration using information about the driving environment and with the expectation that the *human driver* perform all remaining aspects of the *dynamic driving task*	Human driver and system	Human driver	Human driver	Some driving modes
2	**Partial Automation**	the *driving mode*-specific execution by one or more driver assistance systems of both steering and acceleration/ deceleration using information about the driving environment and with the expectation that the *human driver* perform all remaining aspects of the *dynamic driving task*	**System**	Human driver	Human driver	Some driving modes
Automated driving system **("system") monitors the driving environment**						
3	**Conditional Automation**	the *driving mode*-specific performance by an *automated driving system* of all aspects of the dynamic driving task with the expectation that the *human driver* will respond appropriately to a *request to intervene*	System	**System**	Human driver	Some driving modes
4	**High Automation**	the *driving mode*-specific performance by an automated driving system of all aspects of the *dynamic driving task*, even if a *human driver* does not respond appropriately to a *request to intervene*	System	System	**System**	Some driving modes
5	**Full Automation**	the full-time performance by an *automated driving system* of all aspects of the *dynamic driving task* under all roadway and environmental conditions that can be managed by a *human driver*	System	System	System	**All driving modes**

Copyright © 2014 SAE International. The summary table may be freely copied and distributed provided SAE International and J3016 are acknowledged as the source and must be reproduced AS-IS.

Figure 2.2: SAE infographic describing the six levels of automation, with greater technical detail about the definitions.

2.1 *Automation Levels Describe Vehicle Features*

Most new vehicles today have a collection of Level 1 features, such as *automated emergency braking*, *traction control*, and *cruise control*.

There are some high-end vehicles for sale to North American consumers today that have Level 2 automation features, which are usually in the form of assisted highway driving, such as GM's Super-Cruise or Tesla's AutoPilot.

There are companies in North America offering taxi-like services to the public with experimental vehicles that have Level 3 or Level 4 automation. These vehicles are not for sale to the public. They usually have a human safety driver at the wheel, just in case.

A vehicle with some Level 3 functionality might be referred to as a Level 3 vehicle, but it is really the feature that is Level 3. That vehicle is also highly likely to have a collection of Level 1 and 2 features.

2.2 *Models of Human-Machine Collaboration*

The SAE J3016 standard describes a linear scale leading towards *full automation*. This ranking suggests that 'more is better.' In some contexts, it might be more productive to think about different models of human-machine collaboration without being distracted by this linear ranking.

2.2.1 *Assistant: Human Supervision of Machine Control*

Machines are good at repetitive tasks, where there is little or no variation between repetitions. AI technologies do not change this essential truth: they just expand the set of tasks that machines can do, and increase the degree of variation between repetitions that they can cope with. One model of human+machine collaboration is that the machine does the repetitions that fall within a narrow range of variation, and the human handles the more complex cases.

Ideally, the machine can identify which cases are within its capability, so it does not attempt to perform repetitions it is not properly capable of. In the context of automotive, aerospace, and offroad vehicles (the three industry segments that SAE represents), this *hand-off* of control from machine to human can be challenging, because the vehicle often requires timely control actions to continue driving safely.

GM's SuperCruise feature (available on select Cadillac models) can keep the car centred in its lane and at an appropriate following distance from the vehicle in front of it on designated highways, so the human driver does not need to operate the steering wheel nor the pedals if the environment is calm and predictable. Nevertheless, SuperCruise still requires that the driver is responsible for monitoring the driving environment, and for assuming control if an anomalous situation arises. So SuperCruise is similar to the auto-pilot features found on commercial aircraft, where the pilot is still responsible for monitoring the environment and the machine at all times.

Humans can easily become bored and distracted and unfocused when tasked with monitoring a machine performing routine operations in a calm and predictable environment. Airlines address these concerns primarily by training and paying pilots, and through the immense social responsibility of captaining a ship with passengers, but also through a variety of on-the-job strategies. GM's SuperCruise addresses these concerns with a system that tracks the driver's eyeballs to make sure that they are looking at the road, and so at least marginally engaged in monitoring the driving environment. If the driver stops looking at the road, then the car will beep, vibrate the seat and steering wheel, *etc.* If the driver does not respond to these auditory and haptic prompts, the car will will attempt to pull over and phone GM's OnStar service for a human to speak with the driver.

2.2.2 *Guardian: Machine Supervision of Human Control*

Machines can have faster reaction times than humans, and so in some contexts it can make sense for the machine to take over control. A common automotive example is *automated emergency braking:* when the vehicle detects that collision is imminent, it applies the brakes with greater speed and force than the driver is capable of; it might also pre-tension the seatbelts and take other protective actions to prepare the cabin for occupant impact. Here, control is asserted by either the human or the machine, but not both simultaneously.

2.2.3 *Blended: Simultaneous Control*

Dr Gill Pratt of Toyota described their work on developing a system with *blended* human+machine control, as is done in fighter jets, in his remarks at the *Consumer Electronics Show 2019*:

https://global.toyota/en/newsroom/corporate/26085202.html

> With Guardian, the driver is in control of the car at all times except in those cases where Guardian anticipates a pending incident, alerts the driver and decides to employ a corrective response in coordination with driver input.
>
> In this way, Guardian combines and coordinates the skills and strengths of the human and the machine.
>
> In fact, one of the most significant advancements we've made to Guardian this year was the creation of blended envelope control of both the human and the machine. We're inspired and informed by the way that modern fighter jets operate where you have a pilot that flies the stick, but actually they don't fly the plane directly.
>
> Instead, the pilot's intent is translated by a low-level flight control system, to stabilize the aircraft and stay within a safety envelope
>
> So, what does it feel like to have blended envelope control in a car? Most of the time, the driver feels 100 percent in control of the car.
>
> However, as the driver begins to reach the edge of a dynamically changing safety envelope, the machine begins to collaborate with the human driver, nudging the driver back into a safe corridor.
>
> Now, the big idea we really want to drive home is that envelope control is not a discrete on-off switch between the human and the machine.
>
> Rather, it's a seamless blend of both, human and machine working together as teammates where we extract the best skills from each one.

2.2.4 *Partner: Shared Aspects of Control*

It is possible for human and machine to work together, each exercising different aspects of control. For example, with *adaptive cruise control*, the machine controls acceleration and following distance (braking) while the human steers. Conversely, with *lane centering*, the machine steers while the human controls acceleration and braking.

These shared aspects of control are different than the blended control described above: in the blended control, human and machine are working together on a single aspect of control, whereas here each has an independent aspect of control.

The methodology of this report is discussed above in §1. The basic idea is to take NHTSA's text and structure and substitute *advanced financial system* (AFS) where NHTSA's text says *advanced driving system* (ADS), remove text about the physicality of vehicles, and add some finance-specific ideas from UW professors as discussed at the UW+OSFI Workshop. OSFI had no input into this report: it is an academic opinion.

3 Adaptation of NHTSA's Proposed Voluntary Guidance for Autonomous Vehicles

3.1 System Safety

FI are encouraged to adopt voluntary guidance, best practices, design principles, and standards developed by established and accredited standards-developing organizations (as applicable), as well as standards and processes available from other industries.

The design and validation process should also consider including a hazard analysis and safety risk assessment for the AFS, for the overall FI in to which it is being integrated, and when applicable, for the broader financial ecosystem.

Additionally, the process should describe design redundancies and safety strategies for handling AFS malfunctions.

The process should place significant emphasis on software development, verification, and validation. The software development process is one that should be well-planned, well-controlled, and well-documented to detect and correct unexpected results from software updates. Thorough and measurable software testing should complement a structured and documented software development and change management process and should be part of each software version release.

Design decisions should be linked to the assessed risks that could impact safety-critical system functionality.

Design safety considerations should include design architecture, potential software errors, reliability, and bias.

See discussions of *bias* in §3.5.1 and §5.3. Also, discussions of *explainability* in §3.5.2 and §5.4.

All design decisions should be tested, validated, and verified as individual subsystems and as part of the entire system architecture.

FI are encouraged to document the entire process: all actions, changes, design choices, analyses, associated testing, and data should be traceable and transparent.

3.2 *Operational Design Domain (ODD)*

FI are encouraged to define and document the Operational Design Domain (ODD) for each AFS as tested or deployed, as well as document the process and procedure for assessment, testing, and validation of AFS functionality with the prescribed ODD. The ODD should describe the specific conditions under which a given AFS or feature is intended to function.

FI operate in a societal context, and so entities are encouraged to draw on theories and concepts from the social sciences and humanities in defining the ODD of their AFS, as well as describing financial, technical, and logistical aspects of the ODD.

Inspired by Prof Anindya Sen's comments at the UW+OSFI Workshop.

The training data used in the creation of the AFS is also an important aspect of its ODD, and should be characterized and archived. FI are encouraged to have a review schedule to ensure that the ODD reflects market conditions, and to have an update or phase-out plan for when the ODD no longer reflects market conditions.

Inspired by Prof Alexander Wong's comments at the UW+OSFI Workshop.

An AFS should be able to operate safely within the ODD for which it is designed. In situations where the AFS is outside of its defined ODD or in which conditions dynamically change to fall outside of the AFS's ODD, the system should transition to a minimal risk condition, which could entail transitioning control to a receptive, fallback-ready human operator, or to another lower-risk system. In cases the AFS does not have indications that the operator is receptive and fallback-ready, the system should continue to mitigate manageable risks.

To support the safe introduction of AFS in the financial system and to speed deployment, the ODD concept provides the flexibility for FI to initially limit the complexity of broader challenges in a confined ODD.

3.3 *Event Detection and Response*

Event Detection and Response (EDR) refers to the detection by the FI or AFS of any circumstance that is relevant to the immediate decision making task, as well as the implementation of the appropriate FI or system response to such circumstance. For the purposes of this Guidance, an AFS is responsible for performing EDR while it is engaged and operating in its defined ODD. Entities are encouraged to have a documented process for assessment, testing, and validation of their AFS's EDR capabilities. When operating within its ODD, an AFS's EDR functions are expected to be able to detect and respond to events that could affect safe operation of the financial system. An AFS's EDR should also include the ability to address a wide variety of foreseeable encounters, including emergency and other unusual conditions that may impact the safe operation of an AFS.

Normal operations. FI are encouraged to have a documented process for the assessment, testing, and validation of a variety of behavioral competencies for their AFS. Behavioral competency refers to the ability of an AFS to operate in the market conditions that it will regularly encounter, including obeying financial laws and regulations, and responding to hazards. The full complement of behavioral competencies a particular AFS would be expected to demonstrate and routinely perform will depend upon the individual AFS, its ODD, and the designated fallback (minimal risk condition) method. FI are encouraged to consider all known behavioral competencies in the design, test, and validation of their AFS.

Crash Avoidance Capability — Hazards. FI are encouraged to have a documented process for assessment, testing, and validation of their crash avoidance capabilities and design choices. Based on the ODD, an AFS should be able to address applicable pre-crash scenarios that relate to control loss. Depending on the ODD, an AFS may be expected to handle many of the pre-crash scenarios that ~~OSFI has~~ identified previously.

NHTSA's concern is crashes of individual vehicles, whereas OSFI is likely concerned with crashes of entire FI or markets. What exactly 'crash' and 'safety' and 'hazard' mean in finance is left as an exercise for the reader, as discussed in §1.4.1.

In real life, OSFI has done no such thing: to our knowledge they have never done anything like identifying a pre-crash scenario. Remember that this is NHTSA's text, and we have just changed 'NHTSA' to 'OSFI'.

3.4 Fallback (Minimal Risk Condition)

FI are encouraged to have a documented process for transitioning to a minimal risk condition when a problem is encountered or the AFS cannot operate safely. An AFS should be capable of detecting that the AFS has malfunctioned, is operating in a degraded state, or is operating outside of the ODD. Furthermore, an AFS should be able to notify a human operator of such events in a way that enables the attendant to regain proper control of the AFS or allows the AFS to return to a minimal risk condition independently.

Fallback actions are encouraged to be administered in a manner that will facilitate safe operation of the AFS and minimize erratic behaviour. Such fallback actions should also consider minimizing the effects of errors in human recognition and decision-making during and after transition to manual control. In cases of higher automation in which a human operator may not be available, the AFS must be able to fallback into a minimal risk condition without the need for human intervention.

A minimal risk condition will vary according to the type and extent of a given failure, but may include automatically bringing the AFS to a safe stop, preferably. FI are encouraged to have a documented process for assessment, testing, and validation of their fallback approaches.

3.5 *Validation Methods*

Given that the scope, technology, and capabilities vary widely for different automation functions, FI are encouraged to develop validation methods to appropriately mitigate the safety risks associated with their AFS approach. Tests should demonstrate the behavioural competencies an AFS would be expected to perform during normal operation, the AFS's performance during failure avoidance situations, and the performance of fallback strategies relevant to the AFS's ODD.

To demonstrate the expected performance of an AFS for public deployment, test approaches may include a combination of simulation, test runs with historical data, and human-monitored execution in live market conditions.

Prior to execution in live market conditions, FI are encouraged to consider the extent to which simulation and historical-data testing may be necessary. Testing may be performed by the FI themselves, but could also be performed by an independent third party.

FI should continue working with OSFI and other regulatory bodies, industry standards organizations, professional organizations, and others to develop and update tests that use innovative methods as well as to develop performance criteria for test facilities that intend to conduct validation tests.

3.5.1 *Bias & Risk*

This subsection does not appear in NHTSA's guidance.

AFS should not be *biased* in their decisions based on inputs that are protected by the Canadian Charter of Rights and Freedoms or the various provincial Human Rights Codes, *etc.* FI should provide evidence that their AFS have not been trained on protected attributes.

Bias was a recurring theme in the UW+OSFI Workshop discussions, and is discussed further in §5.3.

Dataset augmentation can be an effective technique for improving machine learning, however the augmentation is only as effective as the model on which the augmentation is based. In the case of financial data, dataset augmentation could lead to biases or mis-calculated risk, depending on how the augmented samples are generated. When dataset augmentation is used to develop an AFS, then FI should provide evidence that the augmentation model is not based on protected attributes (*e.g.*, the Canadian Charter of Rights and Freedoms prohibits discrimination on race), and that it does not expose the FI or its customers or the financial ecosystem to mis-calculated risks.

Inspired Prof Paul Fieguth's comments at the UW+OSFI Workshop.

3.5.2 Explainability

This subsection does not appear in NHTSA's guidance.

AFS should be *explainable:* that is, a human should be able to understand why the AFS made its decisions.

Explainability was a general theme at the UW+OSFI Workshop, and is discussed further in §5.4.

Explanations might take various forms, such as a chain of logic, or a highlighting of certain aspects of the input that were most important for the decision, or comparison to a simpler system, or perhaps other forms.

Prof Alex Wong introduced the idea of highlighting part of the input as a form of explanation.

Comparative assessment uses a simpler and more understandable system to validate a more complex one. Both systems compute essentially the same function, but the more complex system might have more desirable outputs. The simpler system can serve to explain and validate the more complex system. For example, a *support vector machine* (SVM) might be used to explain and validate a *neural network*.

Inspired by Prof Paul Fieguth's comments at the UW+OSFI Workshop.

3.6 Human Machine Interface

Understanding the interaction between the system and the operator, commonly referred to as "human machine interface" (hmi), can play an important role in the financial software design process. New complexity is introduced to this interaction as afs take on decision-making functions, in part because in some cases the system must be capable of accurately conveying information to the human operator regarding intentions and system performance. This is particularly true for afs in which human operators may be requested to perform any part of the decision making process in a timely manner.

For example, in a Level 3 afs, the operator always must be receptive to a request by the system to take back decision-making responsibilities. However, an operator's ability to do so is limited by their capacity to stay alert to the operating task and thus capable of quickly taking over control, while at the same time not performing the actual decision-making task until prompted by the system. fi are encouraged to consider whether it is reasonable and appropriate to incorporate operator engagement monitoring in cases where operators could be involved in the decision-making task so as to assess operator awareness and readiness to perform the full decision-making task.

nhtsa's text is written from the perspective of an individual driver having to take control of an individual vehicle in the case where an ads requests it — which is known to be difficult for people to do effectively.

fi are also encouraged to consider and document a process for the assessment, testing, and validation of the system's hmi design. Considerations should be made for the human operator, occupant(s), and external actors with whom the afs may have interactions, including other systems (both traditional and those with afs). hmi design should also consider the need to communicate information regarding the afs 's state of operation relevant to the various interactions it may encounter and how this information should be communicated.

In systems that are anticipated not to have operator controls, entities are encouraged to design their hmi to accommodate people with disabilities (*e.g.*, through visual, auditory, and haptic displays).

In systems where an afs may be intended to operate without a human operator or even any human supervisor, the remote dispatcher or central control authority, if such an entity exists, should be able to know the status of the afs at all times. Examples of these may include Automation Level 4 or 5 systems.

Given the ongoing research and rapidly evolving nature of this field, fi are encouraged to consider and apply voluntary guidance, best practices, and design principles published by appropriate standards bodies and professional organizations, based upon the level of automation and expected level of operator engagement.

3.7 *Cybersecurity*

FI are encouraged to follow a robust product development process based on a systems engineering approach to minimize risks to safety, including those due to cybersecurity threats and vulnerabilities. This process should include a systematic and ongoing safety risk assessment for each AFS, the overall financial product design into which it is being integrated, and when applicable, the broader financial ecosystem.

FI are encouraged to design their AFS following established best practices for financial systems. FI are encouraged to consider and incorporate voluntary guidance, best practices, and design principles published by relevant standards bodies and professional organizations, as appropriate. ~~OSFI encourages~~ FI to document how they incorporated financial cybersecurity considerations into AFS, including all actions, changes, design choices, analyses, and associated testing, and ensure that data is traceable within a robust document version control environment.

In real life, OSFI has not, to our knowledge, issued any guidance on this. Recall that this is NHTSA's text, and we have just changed 'NHTSA' to 'OSFI', as described in §1.

Financial systems are increasingly complex, with a growing number of advanced, integrated functions. Financial systems are also more reliant than ever on multiple paths of connectivity to communicate and exchange data, and they depend on commodity technologies to achieve functional, cost, and marketing objectives. FI encompass a broad sector of the economy and require coordination across all levels of government and the private sector in the event of a significant cyber incident to enable shared situational awareness and allow for a unified approach to sector engagement.

Industry sharing of information on financial cybersecurity facilitates collaborative learning and helps prevent industry members from experiencing the same cyber vulnerabilities. FI are encouraged to report to the Financial Services Information Sharing and Analysis Center (FS-ISAC) all discovered incidents, exploits, threats and vulnerabilities from internal testing, consumer reporting, or external security research as soon as possible, regardless of membership. FI are further encouraged to establish robust cyber incident response plans and employ a systems engineering approach that considers financial system cybersecurity in the design process. FI involved with AFS should also consider adopting a coordinated vulnerability reporting/disclosure policy.

3.8 Crashworthiness

Consumer Protection. Given that a mix of fi with afs and those without will be operating in the public financial system for an extended period of time, fi still need to consider the possible scenario of another fi crashing into an afs-equipped fi and how to best protect consumers in that situation. Regardless of whether the afs or a human operator is in charge of the decision-making task, the consumer protection system should maintain its intended performance level in the event of a crash.

nhtsa's text here obviously focuses on 'crashes' involving individual vehicles and their occupants. What exactly 'crashworthiness' means in the financial context is left as an exercise for the reader, as discussed in §1.4.1.

Entities should consider incorporating information from the advanced sensing technologies needed for afs operation into new consumer protection systems that provide enhanced protection to consumers of all ages and financial means.

In engineering terms, this concept is sometimes known as a *runtime safety monitor:* a separate system that is simpler than the afs, and which monitors the afs performance.

3.9 Post-Failure AFS Behaviour

fi engaging in testing or deployment should consider methods of returning afs to a safe state immediately after being involved in a failure. Depending upon the severity of the failure, actions such as powering off the system, freezing controls, moving the system to a safe state, and other actions that would assist the afs should be considered. If communications with first responders, government agencies, or failure notifications exist, relevant data is encouraged to be communicated and shared to help reduce the harm resulting from the failure.

Additionally, fi are encouraged to have documentation available that facilitates the maintenance and repair of afs before they can be put back in service. Such documentation would likely identify the equipment and the processes necessary to ensure safe operation of the afs after repairs.

3.10 Data Recording

See discussion in §5.5.

Learning from failure data is a central component to the safety potential of AFS. For example, the analysis of a failure involving a single AFS could lead to safety developments and subsequent prevention of that failure scenario in other AFS. Paramount to this type of learning is proper failure reconstruction. Currently, no standard data elements exist for law enforcement, researchers, and others to use in determining why an AFS failed. Therefore, FI engaging in testing or deployment are encouraged to establish a documented process for testing, validating, and collecting necessary data related to the occurrence of malfunctions, degradations, or failures in a way that can be used to establish the cause of any failure. Data should be collected for testing and use, and FI are encouraged to adopt voluntary guidance, best practices, design principles, and standards issued by accredited standards developing organizations. Likewise, these organizations are encouraged to be actively engaged in the discussion and regularly update standards as necessary and appropriate.

To promote a continual learning environment, FI engaging in testing or deployment should collect data associated with failures involving: (1) financial harm to members of the public, or (2) disengagement of the AFS from its task.

What kind of failures should require data collection? NHTSA's text says:
(1) fatal or nonfatal personal injury or
(2) damage that requires towing.

For failure reconstruction purposes (including during testing), it is recommended that AFS data be stored, maintained, and readily available for retrieval as is current practice, including applicable privacy protections, for failure event data recorders. Systems should record, at a minimum, all available information relevant to the failure, so that the circumstances of the failure can be reconstructed. These data should also contain the status of the AFS and whether the AFS or the human operator was in control of the system leading up to, during, and immediately following a failure. FI should have the technical and legal capability to share with government authorities the relevant recorded information as necessary for failure reconstruction purposes. Meanwhile, for consistency and to build public trust and acceptance, ~~OSFI will continue~~ working with appropriate standards bodies and professional organizations to begin the work necessary to establish uniform data elements for AFS failure reconstruction.

In real life, we do not know what OSFI is working on. Recall that this is NHTSA's text, and we have just changed 'NHTSA' to 'OSFI', as described in §1.

For some kinds of AFS, failures are often rooted in incomplete or biased training data, and so training data should also be archived and accessible, as appropriate.

Inspired by multiple discussions at the UW/OSFI workshop.

3.11 Consumer Education and Training

Education and training is imperative for increased safety during the deployment of AFS. Therefore, FI are encouraged to develop, document, and maintain employee, dealer, distributor, and consumer education and training programs to address the anticipated differences in the use and operation of AFS from those of the conventional systems that are owned and operated on today. Such programs should consider providing target users the necessary level of understanding to utilize these technologies properly, efficiently, and in the safest manner possible. FI, particularly those engaging in testing or deployment, should also ensure that their own staff, including their marketing and sales forces, understand the technology and can educate and train their dealers, distributors, and consumers.

Consumer education programs are encouraged to cover topics such as AFS ′ functional intent, operational parameters, system capabilities and limitations, engagement/disengagement methods, HMI, emergency fallback scenarios, operational design domain parameters (*i.e.*, limitations), and mechanisms that could alter AFS behaviour while in service. They should also include explicit information on what the AFS is capable and not capable of in an effort to minimize potential risks from user system abuse or misunderstanding.

As part of their education and training programs, AFS dealers and distributors should consider including a demonstration of AFS operations and HMI functions in an appropriately realistic environment prior to consumer release. Other innovative approaches may also be considered, tested, and employed. These programs should be continually evaluated for their effectiveness and updated on a routine basis, incorporating feedback from dealers, customers, and other sources.

3.12 Federal, Provincial, and Local Laws

FI are also encouraged to document how they intend to account for all applicable federal, provincial, and local laws in the design of their AFS. Based on the ODD, the development of AFS should account for all governing financial regulations and consumer protection laws when operating in automated mode for the region of operation.
For testing purposes, a FI may rely on an AFS test operator or other mechanism to manage compliance with the applicable laws.

In certain safety-critical situations human operators may temporarily violate certain regulations. It is expected that AFS have the capability of handling such foreseeable events safely; FI are encouraged to have a documented process for independent assessment, testing, and validation of such plausible scenarios.

Given that laws and regulations will inevitably change over time, FI should consider developing processes to update and adapt AFS to address new or revised legal requirements.

4
Professional Organizations, Licensure, and Ethics

4.1 Professional Organizations & Standards Bodies

Professional organizations and standards bodies play important complementary role to regulators in guiding the practice of professionals. This is true in engineering as it is in other professional fields. In the specific example of autonomous vehicles, one can observe the complementary roles of SAE and NHTSA.

There are a wide range of professionals already involved in Canada's financial system, and their existing professional organizations and standards bodies have important roles to play in guiding the prudent use of AI technologies in this context.

Engineers and computer scientists might play increasingly important roles in Canada's financial system, and so it is prudent to also follow the guidelines of their professional organizations and standards bodies, as relevant and appropriate.

In Canada, the practice of professional engineering is regulated at the provincial level. Each province has its own legislation and corresponding professional organization. For example, the University of Waterloo is in the province of Ontario, and so the relevant professional body is *Professional Engineers Ontario* (PEO).

https://www.peo.on.ca/

Internationally, the two most important professional organizations related to computing are the *Association for Computing Machinery* (ACM) and the *Institute of Electrical and Electronics Engineers* (IEEE). The ACM was founded in 1947, and is the world's largest scientific and educational computing society, with almost 100,000 members around the world. The IEEE is the world's largest technical professional organization dedicated to advancing technology for the benefit of humanity, with over 400,000 members around the world. The ACM and the IEEE collaborate on a number of important topics of shared interest, such as educational curriculum recommendations and professional ethics.

https://www.acm.org

https://www.ieee.org/

There are a number of new organizations around AI specifically, and their recommendations are surely worth heeding as well.

4.2 *Professional Licensure*

Some professional organizations in some jurisdictions have specific powers of licensing and disciplining practitioners. So it is with the provincial professional engineering organizations in Canada. This is not the case for the international bodies of the ACM and the IEEE.

In the more established branches of engineering, a professional license is required to practice. For example, the design and construction of a bridge must be overseen by a licensed engineer.

The purpose of professional licensure and discipline is to protect the public interest and maintain the public trust. It is prudent to protect these concerns in Canada's financial system, just as it is for Canada's physical infrastructure.

4.2.1 *Recommendation*

The reader might expect that we recommend licensed engineers in positions of oversight of AI technologies within Canada's financial system, but we do not make this recommendation for four reasons:

1. Canada's financial system already involves a wide range of professionals from other disciplines, and in specific contexts they might be more appropriate overseers.

2. Professional licensure is uncommon for those educated as software engineers. Of the over 1,000 students who have graduated from the Software Engineering program at UW, slightly less than 1% have gone on to become licensed by PEO.

3. Professional licensure is difficult to obtain for those educated as software engineers. Licensure typically requires several years of experience supervised by a licensed engineer. Our graduates tell us that it is rare for them to have a supervisor who is licensed, so it is difficult to have their experience count towards licensure.

4. Many of the people with the appropriate technical skill to design and deploy AI technologies will have been educated in computer science programs rather than software engineering programs. They might have all of the requisite mathematical and computing knowledge for working with AI technologies, but lack the education in physics and chemistry that is required to become a licensed engineer in Canada.

We recommend that AFS development and deployment be overseen by someone who is a member of a relevant professional organization. This recommendation includes the many varieties of financial professionals that have historically worked in Canada's FI. In the technical realm, it also includes provincial bodies such as the PEO, as well as international bodies such as the ACM and IEEE.

4.3 Professional Ethics

Many professional organizations have a code of ethics, and this is the case for the ACM, the IEEE, and Canada's provincial engineering organizations. Almost all of these codes begin with an acknowledgement of the public interest. For example:

Joint ACM/IEEE Code of Ethics for Software Engineers: 1. PUBLIC — Software engineers shall act consistently with the public interest.

https://ethics.acm.org/code-of-ethics/software-engineering-code/

https://www.computer.org/education/code-of-ethics

ACM General Code of Ethics: 1.1 Contribute to society and to human well-being, acknowledging that all people are stakeholders in computing.

http://ethics.acm.org/code-of-ethics/

IEEE Code of Ethics: We, the members of the IEEE, in recognition of the importance of our technologies in affecting the quality of life throughout the world, and in accepting a personal obligation to our profession, its members, and the communities we serve, do hereby commit ourselves to the highest ethical and professional conduct and agree:

https://www.ieee.org/about/corporate/governance/p7-8.html

1. to hold paramount the safety, health, and welfare of the public, to strive to comply with ethical design and sustainable development practices, and to disclose promptly factors that might endanger the public or the environment;

PEO Code of Ethics:

https://www.ontario.ca/laws/regulation/900941

Every province's engineering organization has a code of ethics. PEO is chosen as an example here because it is the provincial organization that we teach about at the University of Waterloo.

1. It is the duty of a practitioner to the public, to the practitioner's employer, to the practitioner's clients, to other members of the practitioner's profession, and to the practitioner to act at all times with,
   - i fairness and loyalty to the practitioner's associates, employer, clients, subordinates and employees,
   - ii fidelity to public needs,
   - iii devotion to high ideals of personal honour and professional integrity,
   - iv knowledge of developments in the area of professional engineering relevant to any services that are undertaken, and
   - v competence in the performance of any professional engineering services that are undertaken.
2. A practitioner shall,
   - i regard the practitioner's duty to public welfare as paramount,

The IEEE *Global Initiative on Ethics of Autonomous and Intelligent Systems.* From the preamble of their report:

https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

- The IEEE Global Initiative's mission is, "To ensure every stakeholder involved in the design and development of autonomous and intelligent systems is educated, trained, and empowered to prioritize ethical considerations so that these technologies are advanced for the benefit of humanity."
- Whether our ethical practices are Western (*e.g.*, Aristotelian, Kantian), Eastern (*e.g.*, Shinto, School of Mo, Confucian), African (*e.g.*, Ubuntu), or from another tradition, honoring holistic definitions of societal prosperity is essential versus pursuing one-dimensional goals of increased productivity or gross domestic product (GDP). Autonomous and intelligent systems should prioritize and have as their goal the explicit honoring of our inalienable fundamental rights and dignity as well as the increase of human flourishing and environmental sustainability.

4.3.1 Recommendations

The stability and trustworthiness of Canada's financial system is vital to the public interest. The codes of ethics from most of the relevant technical professional organizations emphasize the importance of protecting the public interest. Therefore:

1. We recommend that computing professionals working within Canada's financial system belong to an established professional organization with a code of ethics.
2. We suggest, for consideration, that professionals post the code of ethics from whichever organization(s) they belong to in a prominent location in the workplace.

This suggestion is inspired by the PEO code of ethics, which has a clause requiring engineers to post their licences in a prominent location in the workplace.

5 Discussion

The core of this report is an adaptation of NHTSA's *Proposed Voluntary Guidance for Autonomous Vehicles v2.0* to the context of Canada's financial system. A caricature of this adaptation would be that we just changed NHTSA's *advanced driving system* (ADS) to *advanced financial system* (AFS) — then we sprinkled in some ideas that were discussed at the UW+OSFI Workshop.

This adaptation seems to have worked quite well, in the sense that it is our opinion as academics and engineers that Canada's financial institutions (FI) should be following all of this guidance. Maybe there are more guidelines that Canada's FI should also be following in the context of their use of AI, but this seems to us to be a reasonable baseline starting position.

Why does this adaptation work so well? There are several reasons. Despite the obvious differences between autonomous vehicles and banks and insurance companies, there are two critical similarities: first, both transportation and finance are essential for modern society; second, both involve complex processes that are currently performed by people with significant mechanical assistance, and the discussion is largely about automating these decision-making processes.

An important reason why this adaptation works so well is that much of NHTSA's language describes professional engineering practice, which is good accumulated wisdom applicable in any area where complex technological systems are designed and deployed. We expect that some of this engineering guidance is already practiced by Canada's financial institutions: for example, following a documented process. But there are some engineering concepts in NHTSA's report that might not be as well known in Canada's FI. For example, *operational design domain* is a foundational concept in NHTSA's guidance that is used when discussing a variety of issues, which are summarized below.

ODD is introduced in §3.2.

Bias and *explainability* are two concepts of great concern in many discussions of AI, and were central themes of discussion at the

UW+OSFI Workshop, but do not feature in NHTSA's language. We added sections §3.5.1 and §3.5.2 on these topics above, and discuss them further below.

Finally, this discussion concludes with a consideration of the ends to which AI might be used, informed by NHTSA's position on this issue in the context of autonomous vehicles.

5.1 Operational Design Domain

Snow tires are designed for driving in snow. Racing tires are designed for driving race cars on race tracks. Tractors use tires specially designed for farming. And so on. There are many different kinds of tires for many different *operational design domains* (ODD). There are some particular kinds of technical products, such as tires, that are generally understood to be specialized to different ODD.

In product categories less richly developed than tires, it might be less obvious that each technical product still has an ODD. Characterizing the ODD is an important part of every engineered system: this is the range of parameters within which the design has been analyzed and verified, and within which it is intended to perform well.

This report, following NHTSA's structure and language, introduces the ODD concept in §3.2, and then uses that concept in discussing a variety of other important issues. A summary of the ODD-related recommendations:

1. Every AFS should have a defined ODD. This definition should probably include at least financial, technical, and logistical aspects.

2. Because FI operate in a societal context, these ODD definitions should be informed by appropriate theories and concepts from the humanities and social sciences.

Inspired by Prof Anindya Sen's comments at the UW+OSFI Workshop.

3. When operating within its ODD, and AFS is expected to be able to detect and respond to events that could affect its safe operation.

Discussed in §3.3.

4. An AFS should have mechanisms in place to detect if it is being operated outside of its ODD. If this occurs, the AFS should notify human operators and fallback to a minimal risk condition.

Discussed in §3.4.

5. The training data used in the creation of the AFS is also an important aspect of its ODD, and should be characterized and archived. FI are encouraged to have a review schedule to ensure that the ODD reflects market conditions, and to have an update or phase-out plan for when the ODD no longer reflects market conditions.

Inspired by Prof Alexander Wong's comments at the UW+OSFI Workshop.

5.2 *Human-Machine Interaction*

Discussion of §3.6.

The public imagination is captivated by the idea of super-human AI technologies, as portrayed in books and movies such as *Neuromancer*, *2001: A Space Odyssey*, and *Terminator*. These super-human AI technologies are entirely autonomous and independent of human supervision or control, and furthermore they are not constrained to a well-defined ODD.

It's hard to know exactly how technology will evolve in the future, but such super-human AI technologies are not likely in the near term. Where technology is at today, and what is likely that we will see more of in the coming years, are advanced automation systems with narrowly-defined ODD and with significant human-machine interaction. Even if the automation system takes over control of operations for a longer time duration than is currently the case, the human-machine interaction will still exist and will need perhaps even more sophisticated and careful design.

The case of commercial airplane pilots might be instructive. Commercial jetliners have had autopilot features for decades. Nevertheless, commercial flights still have highly trained human pilots in the cockpit. Their training involves not only how to fly the plane manually, but how to respond when the automation produces undesirable behaviours in flight. So while the automation has reduced the amount of work that the pilots do during the flight, it has required greater training to become a pilot.

Similarly, the SAE AutoDrive Challenge requires the autonomous vehicle safety drivers to take training at General Motors in which GM engineers initiate vehicle malfunctions on the test track. For example, unintended acceleration, braking, or steering. Being a safety driver for a student autonomous vehicle team means being prepared for an autonomous vehicle that might not behave as intended or expected.

Another aspect of pilot training might also be instructive: simulations of adverse conditions. Pilots undergo regular training and practice, in simulation, of inclement weather, system malfunctions, crash landings, and other adverse scenarios.

The human-machine interaction is an active concern for all industry segments that SAE represents: automotive, aerospace, and off-road vehicles (*e.g.*, mining and farming). Significant design efforts are invested in the HMI aspects of these systems. Products are differentiated on their HMI features. For example, a distinguishing feature of GM SuperCruise feature (available on select Cadillac models) is that it permits the driver to take their hands off the wheel. This is achieved by a system that monitors the driver's eyeballs, to ensure that they are paying attention to the road.

Recommendation: HMI design for AFS.

HMI design for AFS should be an area of serious investment, development, and research. Human oversight and supervision is still necessary, and the training required to perform those roles should be extensive and include simulations of adverse conditions.

Recommendation: maintain public trust.

Finally, it is important to recognize that maintaining public trust requires highly trained human pilots in the cockpit. Without public trust the commercial air travel business would cease to exist.

5.3 *Bias*

Discussion of §3.5.1.

Bias is a concern in society and in the application of AI technologies to financial decision making. This report has taken the position that AI systems should not be trained on protected attributes. For example, the Canadian Charter of Rights and Freedoms prohibits discrimination on race, gender, sexual orientation, *etc.*

That position might not be strong enough. At the UW+OSFI Workshop, Prof Joel Blit spoke of a study in which the AFS was not trained on any protected attributes, yet still learned patterns that correlated with protected attributes. For example, making decisions about mortgage rates that correlate with race, even though the AFS did not consider race as an input attribute.

However, there are related contexts that appear to be less concerning. For example, at the UW+OSFI Workshop the issue of car insurance for young men was discussed. This group is well known to engage in riskier driving behaviour than many other groups. But in some jurisdictions insurers are prohibited from discriminating based on age. So instead they discriminate based on years of driving experience, which is highly correlated with age. This seems less problematic than having mortgage rates correlate with race.

It is not the place of engineers to make the determinations about which cases of AFS decision-making correlating with protected attributes are socially acceptable and which are not: this is a place where the contributions of thinkers from the humanities and social sciences are vital.

For now it seems that a good first step is to *not* train the AFS on protected attributes, as doing so seems obviously problematic. Engineers should be mindful of the fact that even if their system has not been explicitly trained on protected attributes, its outputs might correlate with protected attributes in ways that do not serve the public interest. Engineers should consult with appropriate others to ensure that their deployed systems perform in the public interest.

5.4 *Explainability*

Brief discussion of §3.5.2.

At the UW+OSFI Workshop, Profs Alex Wong and Paul Fieguth each introduced a specific technical way of establishing explainability for an AFS— and here they specifically had in mind an AFS based on *neural network* technology.

Prof Wong introduced the idea of a neural net reporting which bits of the input were most important in making its decision. This can help a human understand what features the net is and isn't considering when reaching a particular conclusion.

Prof Fieguth introduced the idea of comparing a neural net based system to a simpler and more understandable system as a way to understand the neural network.

5.5 Data

Summary and discussion of data-related points in this report:

- AI *training data* should be considered as part of the system's ODD definition, and should be characterized, archived, and auditable.
- AI systems should have a *periodic review schedule* to ensure that their ODD continues to match current market conditions. In other words, periodically check that the training data corresponds to current data.
- *Dataset augmentation* can be a powerful technique, but can introduce bias and mi-calculated risk if not used carefully.
- *Black boxes:* AI systems should log data as they operate, so that the decisions they have made can be reviewed if necessary. This is also critical for training and re-training the AI systems so that they improve over time.
    - *Aerospace* has more of a tradition of sharing failure data across the industry, so that all may learn collaboratively. Working together to maintain safety and public trust in air travel.
    - *Automotive* has a history of competing on safety: auto makers advertise to the public that their vehicles are safer than competitors. So the automotive industry has less of a culture and history of sharing failure data. Competition, however, pushes the industry collectively forward.
    - Which approach is best for FI: collaborative or competitive?
    - Some jurisdictions, such as California, require public reporting of autonomous vehicle AI disengagement events — not the data from the event, but the existence of the event.

NHTSA's guidance does not say much on the topics of data *privacy* and *ownership*. These are also important topics, and especially so in the financial context.

5.6 *What is the purpose of using AI?*

NHTSA supports the development and deployment of *advanced driving systems* (ADS) because they see potential for these technologies to improve public safety, which is their mandate. All of the codes of ethics from engineering and computing professional organizations reviewed above in §4.3 first emphasize a duty to the public welfare, and public road safety is clearly consistent with that.

While NHTSA is the *National Highway Traffic Safety Administration*, the SAE represents three industry segments: automotive, aerospace, and offroad vehicles. These different industry segments might have varying additional motivations on top of safety.

5.6.1 *Automotive, Aerospace, and Offroad*

AUTOMOTIVE ENGINEERS might additionally be motivated by:

- *Accessability:* an aging population and the differently-abled might have difficulty driving conventional vehicles.
- *Sustainability:* ADS might reduce transportation's carbon footprint, and might also reduce the geographic footprint of vehicle storage.
- *Customer Experience:* While the driving depicted in most TV ads is enjoyable, most consumer miles are driven in unpleasant traffic conditions where ADS could create a better customer experience.

It is less clear how autonomous vehicles create corporate profits, especially in the short term. There is certainly consumer demand today for vehicles with enhanced partial automation features. Fully autonomous vehicles will likely require shifting to new business models, and significant investor capital is chasing these potential future opportunities.

AEROSPACE ENGINEERS might additionally be motivated by:

- *Fuel efficiency:* airline profits are influenced by fuel efficiency, and advanced automation systems might be able to optimize this in a way that human pilots would struggle to replicate.
- *Pilot workload management:* Pilots need to be fully focused and on-task at some very specific parts of a flight, and automation can help manage their workload to achieve those goals.

OFFROAD EQUIPMENT ENGINEERS, such as for farming and mining, might additionally be motivated by assisting human operators in performing work that is dangerous, remote, seasonal, or 24/7. The short-term business case for advanced automation is fairly clear in this industry segment. Companies like John Deere and Caterpillar are active in adding advanced automation to their equipment.

5.6.2 Financial Institutions

From our position as academics and engineers, outside of the world of financial institutions, we imagine that there might be three primary reasons why financial institutions might want to use AI:

1. *To improve profits through shrewder financial decision-making.* All private corporations have a natural interest in improving profits. It is of course in the public interest to have robust financial institutions, and many Canadian pension plans and retail investors are shareholders in Canada's financial institutions.
2. *To improve productivity.* There is much discussion about AI's potential to improve productivity, and perhaps to disrupt parts of the workforce. How this plays in to the public interest is a matter of debate that is larger than the context of Canada's financial institutions, but which is also important in this context.
3. *To reduce risk.* NHTSA's language is very focused on *safety* and *crash avoidance* and *crashworthiness*. These words have obvious meaning in the context of vehicles on the public highways. But perhaps they can also serve as continued inspiration for the tradition of prudent management that Canada's financial institutions have become world-renowned for. Maybe AI technologies can be applied to avoid or mitigate crashes and disruptions in the financial ecosystem.

Many engineers working on autonomous vehicles are personally motivated, as NHTSA and SAE are, by a vision of improved public safety and public welfare. Of course, they are also motivated to build cool new technologies. While the contribution of autonomous vehicles to the public good might be debated from some perspectives, such as labour disruption, there is general consensus within the industry that they will improve public safety. Perhaps there is a similar alignment between the public interest and the application of AI in Canada's financial institutions.