

Autonomy, Operating Domains, and Validation in Robotic Ultrasound: Transferable Lessons and Structural Limits from Autonomous Vehicle Development

Reya Yassir Khawaja

Year 13 | Nord Anglia International School | Dubai, UAE

ABSTRACT

Diagnostic ultrasound is among the most portable and least expensive of the major imaging modalities, yet its reach is constrained by the availability and physical capacity of skilled operators: between 85 and 90 percent of sonographers report work-related musculoskeletal pain, and approximately one in five sustains a career-ending injury [1, 2, 3]. Robotic ultrasound systems have advanced substantially in response and learned acquisition policies now approach expert-level performance on individual anatomical targets [4, 5]. Progress in the field is, however, characterized using autonomy taxonomies inherited from surgical robotics [6, 7, 8]. These are frameworks that describe how much of a task a machine performs while remaining largely silent on the population in which that performance was demonstrated. Across the studies examined in this review, performance metrics are routinely reported, whereas explicit specification of operating conditions is considerably less common.

This review synthesizes the robotic ultrasound and autonomous vehicle safety literatures and then advances three conceptual proposals. The first is a Dual-Axis Autonomy Matrix, which separates acquisition authority from diagnostic authority on the grounds that in ultrasound the sensor is the actuator. This framing identifies a configuration in which a system selects its own evidence and then evaluates it: a failure mode termed evidentiary self-confirmation. The second is an anatomical operational design domain, which unlike an automotive geofence, cannot be observed prior to entry, motivating a pre-scan domain-admission phase. The third is a reporting standard in which out-of-domain refusal rate is treated as a primary reported quantity.

The review further identifies where the analogy fails. Level 3 supervision transfers poorly because a clinician overseeing an autonomous examination has no perceptual channel independent of the machine and therefore inherits its blind spots irrespective of attention. By contrast, applying the zero-failure confidence bound used in autonomous vehicle safety analysis to diagnostic ultrasound indicates that the required exposure is three to five orders of magnitude smaller, so the binding constraint is the cost of adjudicating ground truth rather than the number of examinations.

Implementable proposals follow: shadow-mode data collection, staged expansion under a Predetermined Change Control Plan, and a Retrospective Volume benchmark in which adjudication is performed once and reused.

INTRODUCTION

Diagnostic ultrasound presents a fundamental paradox: while the equipment itself is portable and accessible, its application is constrained by human physical limits and specialized training requirements. Studies estimate that 85 percent to 90 percent of sonographers report chronic pain, with nearly 20 percent sustaining career-ending injuries due to musculoskeletal disorders caused by static loading, shoulder abduction, and wrist extension during image acquisition [1, 2, 3, 9]. As a result, ultrasound availability remains bottlenecked by operator scarcity.

Additionally, operator dependence introduces diagnostic variance. Unlike computerized tomography (CT) or magnetic resonance imaging (MRI), where scan geometry is fixed by the instrument, ultrasound image planes are selected manually. Differences in transducer orientation and applied contact force between operators result in variations in diagnostic measurements, including carotid intima-media thickness, cardiac ejection fraction, and fetal biometry. At present, automated force-controlled robotic systems address these consistency challenges by standardizing probe contact force and execution path [10, 11].

Recent developments in deep reinforcement learning (RL) have further allowed robotic systems to navigate toward standard scan plans automatically [4, 5]. These models are trained on large demonstration datasets and achieve high spatial accuracy and reproducibility, signaling a transition from experimental prototypes toward potential clinical evaluation.

However, current evaluation standards for robotic ultrasound rely primarily on task-completion metrics adapted from surgical robotics frameworks [6, 7, 8]. These metrics solely quantify the degree of automation but do not define the patient populations or physical conditions under which performance was validated. A system evaluated in a narrow cohort may fail when encountering variations in body mass index, tissue depth, acoustic shadowing, or anatomical anomalies.

A useful framework for addressing this gap exists in autonomous vehicle (AV) development. The automotive field transitioned from evaluating basic control tasks to defining Operation Design Domains (ODDs), which outline the explicit bounds covering environmental, structural, and edge-case conditions [12]. Applying ODD concepts to medical robotics offers a structured approach to defining validation boundaries across diverse patient populations.

Concurrently, directly applying autonomous driving frameworks to ultrasound is limited by three structural differences:

- (1) The operating domain is unobservable prior to contact. This means the tissue acoustic properties and subsurface anatomy cannot be assessed until the transducer touches the skin.
- (2) Oversight is inherently dependent. A human supervisor monitoring a remote automated system relies entirely on the image stream generated by the robot itself, limiting their ability to independently evaluate the situation.
- (3) Ultrasound allows for retrospective validation, as acquisition data can be saved and replayed offline against model pipelines to enable repeatable offline benchmarking without exposing patients to risk.

Ultimately, evaluating automated diagnostic ultrasound requires moving beyond basic task-completion rates toward population-level validation frameworks that account for patient diversity and clinical workflow integration.

Section 2: Current State of Robotic Ultrasound

2.1 Contact Regulation and Kinematic Control

Early-generation robotic ultrasound (RUS) systems have largely resolved the problem of maintaining probe contact force against dynamic, deforming tissue under respiratory motion. Current architectures employ hybrid force-motion control schemes to achieve simultaneous position and force regulation, facilitating a soft landing of the transducer onto patient tissue [13, 14]. These frameworks integrate external RGB-D vision sensors for macro-geometry estimation with real-time ultrasound feedback for fine probe adjustments. Quantitative evaluations show that systems with seven degrees of freedom (7-DoF) use force sensing with ultrasound visual servoing to achieve consistency in contact force, image acquisition stability, and downstream metric repeatability compared to both novice and expert clinicians [10]. Consequently, physical contact mechanics and force stabilization no longer constitute primary bottlenecks to clinical translation.

2.2 Autonomous Acquisition Policies and the Generalization Gap

Determining probe placement is a challenge, and two results characterize the difficulty and a demonstrated route past it.

The first is a deep reinforcement learning framework controlling the six-dimensional pose of an ultrasound probe from real-time image feedback, navigating toward standard scan planes of the spine. Incorporating a confidence-based reward that encodes image quality into the learning objective, it reports navigation accuracy of 4.91 mm and 4.65 degrees, with success rates of 92 percent in the intra-patient setting and 46% in the inter-patient setting [4]. The 46-point reduction is informative as it illustrates the challenge of inter-patient generalization. It does not indicate deficiencies in the method, but it quantifies the extent to which an acquisition policy's competence may be bound to the anatomy on which it was trained. Vessel depth, rib spacing, tissue attenuation, and target position relative to external landmarks all vary between individuals, and a policy fitted to one anatomy has learned something narrower than a general scanning capability.

The second result suggests this can be reduced, although at the cost of considerably greater training-data requirements. UltraBot, a fully learning-based autonomous carotid system, trains a unified imitation-learning framework on approximately 247,000 expert demonstrations achieving success rates above 90 percent with up to fivefold gains in reproducibility on diverse, previously unseen populations [5]. The gap, therefore, appears addressable through data volume, at a scale currently accessible to a limited number of institutions and demonstrated for a single anatomical target.

2.3 Path Planning Under Acoustic Constraint

A third strand treats acquisition as planning under physical constraint rather than reactive servoing. Intercoastal imaging illustrates the problem: the liver and heart lie behind a rib cage, opaque to ultrasound, so the reachable imaging volume is determined by the acoustic windows a given anatomy provides. Recent work trains an RL agent in a virtual environment constructed from CT templates with randomly initialized tumors, using a reward shaped to minimize acoustic attenuation and shadowing, and evaluates on unseen patient models [15, 16]. Related systems address automated path planning more generally and encode clinical protocol as in the use of path-based virtual fixtures to standardize the vein-compression maneuver in deep venous thrombosis examination [13].

2.4 Reporting Patterns Across the Studies Examined

These studies, considered together, exhibit a pattern less apparent when each is read individually: performance metrics are routinely reported, whereas explicit definitions of the population and operating conditions over which those claims are intended to hold are considerably less common [4, 5, 11, 16].

A spine navigation study reports inter-patient variance but does not define the population over which its success rate is asserted. The reproducibility comparison was conducted on two phantoms, leaving performance in human anatomy uncharacterized. The intercoastal planning work evaluates on CT-derived virtual models without physical execution. UltraBot, which comes closest, describes its cohort as diverse and previously unseen but does not specify the bounds of the population within which its success rate is claimed to hold. None of these reports how frequently its system encountered conditions under which it should not have proceeded.

A safety analysis reporting a crash-rate reduction without specifying the geographic, environmental, and speed conditions under which the miles were accumulated would be considered incomplete, not because the analysis would be incorrect but because the claims are not falsifiable.

Section 3: Lessons from Autonomous Vehicle Development and Validation

AV development offers an insightful comparison for robotic ultrasound. This is because AV development represents an extensively documented large-scale effort to take a learned perception-action system from laboratory demonstration to supervised public deployment. Furthermore, its challenges are also documented in federal accident investigations and peer-reviewed safety research. This section focuses on the elements of that experience that may be relevant to the development and validation of automated robotic ultrasound systems.

3.1 Levels of Automation Characterize Responsibility Allocation

The framework established by Society of Automotive Engineers (SAE) J3016 outlines six distinct levels of driving automation, ranging from complete manual control to full automation, representing the standard reference for classifying automated functions [12]. This taxonomy describes the distribution of control and oversight between the human user and the automated system during active engagement. As a result, a Level 4 system should not be considered merely as an incremental technical advancement over a Level 3 system. It should instead be treated as a configuration in which the allocation of the driving task and its fallback differ fundamentally. SAE states that the levels are descriptive and technical rather than normative or legal [12]. The standard does not itself assign liability. Because the levels nonetheless specify which agent is responsible for the driving task under given conditions, they have been widely adopted as a reference point in regulatory and liability discussion [6, 7, 8].

3.2 The Operational Design Domain Framework

In automotive engineering, the primary mechanism of outlining system boundaries is the ODD. The ODD defines the specific operating parameters, which include geographic boundaries, infrastructure constraints, environmental conditions, illumination levels, velocity thresholds, and temporal limitations. These are the parameters under which an automated system is certified to operate safely. The ODD converts a broad safety claim into a falsifiable hypothesis. “This system drives safely” cannot be tested. “This system drives safely on mapped surface streets in Chandler, Arizona, in daylight, in dry conditions, below 45 mph” can be tested and can be shown to be false.

Beyond immediate safety validation, the ODD also serves as the foundation for deployment frameworks, enabling incremental domain expansion, standardized benchmarking across various platform architectures, and definition of operational boundaries that need a transfer of control.

3.3 Limitations of Empirical Road Testing for Safety Validation

In 2016, Kalra and Paddock addressed a fundamental question concerning the scale of empirical validation required to prove AV safety relative to human performance. Utilizing a baseline human fatality rate of 1.09 per 100 million vehicle

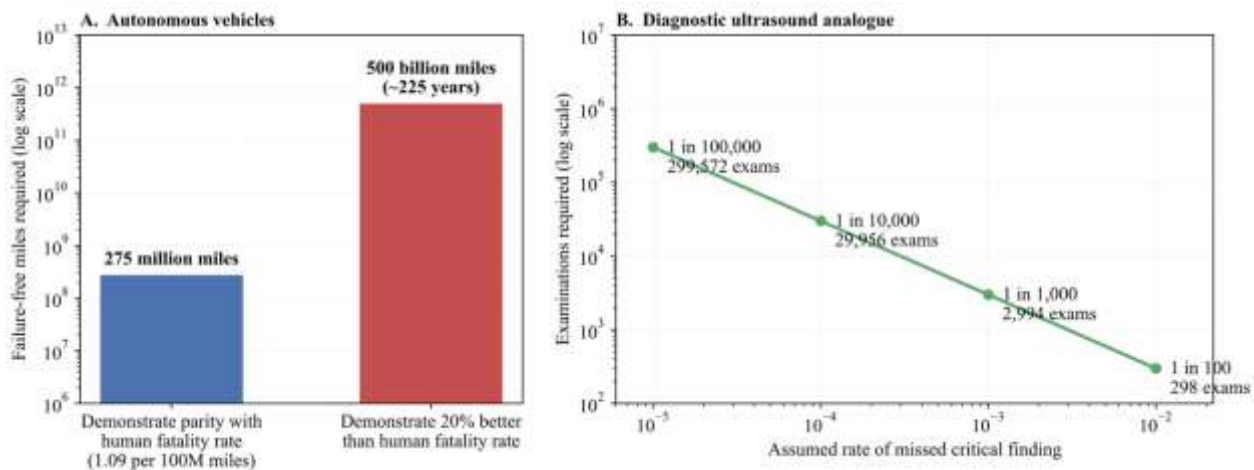
miles traveled, their statistical analysis showed that proving an equivalent fatality rate at a 95% confidence level would require approximately 275 million failure-free miles [17]. Furthermore, demonstrating a 20% reduction in fatality rates would require approximately 500 billion miles of operational testing, a distance equivalent to over two centuries which the authors estimated as roughly 225 years of continuous operation under their stated fleet assumptions. This finding established that direct empirical road exposure is an infeasible pathway for achieving statistical confidence regarding low-frequency, high-severity events. Consequently, the field shifted its validation paradigms toward high-fidelity simulation, structured scenario libraries, staged deployment within strictly defined ODDs, and proxy safety metrics.

An equivalent analysis has not been reported for robotic ultrasound, even though a missed critical diagnostic finding represents a comparably rare, high-consequence failure mode. Figure X applies the same statistical reasoning to both settings. Where R is a failure rate per unit of exposure and n the number of failure-free observations required to place a 95% upper confidence bound on it. $n = \ln(0.05)/\ln(1-R)$. Applied to the human fatality benchmark of 1.09 per 100 million vehicle miles, this yields 274.8 million miles, reproducing the published estimate to within 0.06% and confirming the method. The full derivation is given in Appendix A.

Applied to diagnostic ultrasound, with R interpreted as the rate of missed critical findings per examination, the required exposure is three to five orders of magnitude smaller. Bounding a miss rate below 1 in 1000 requires approximately 3000 failure-free examinations, and below 1 in 10,000 approximately 30,000. These are large but not prohibitive quantities and fall within the scale of a coordinated multi-site study.

Two qualifications apply. The assumed miss rates are illustrative rather than empirical estimates and establishing a defensible target rate remains an open question. The binding constraint in the ultrasound setting is unlikely to be the number of examinations but the cost of establishing ground truth for each one, since determining whether a finding was missed requires an independent reference standard and expert adjudication for every examination in the denominator. The automotive constraint is exposure. The ultrasound constraint is adjudication.

Figure 1. Observations required for a zero-failure safety demonstration at 95% confidence. Both panels apply the relation $n = \ln(0.05) / \ln(1-R)$, which gives the number of failure-free observations needed to place a 95% upper confidence bound on failure rate R . Panel A shows the AV case. Applying the relation to the human benchmark of 1.09 fatalities per 100 million vehicles yields 274.8 million failure-free miles, reproducing the published estimate of approximately 275 million to within 0.06% (Kalra & Paddock, 2016). The 500-billion-mile figure derives from a two-sample comparison against the human benchmark and is reproduced from that source rather than derived here. Panel B applies the same single sample bound to diagnostic ultrasound across a range of assumed missed critical finding rates, with representative values marked. The required number of examinations is three to five orders of magnitude smaller than the automotive requirement. The assumed miss rates are illustrative and are not empirical estimates. Full derivation in Appendix A.



3.4 Disengagement Reporting and the Pitfalls of Safety Metric Design

A regulatory framework established in California mandates that entities testing AVs on public roads submit public disengagement reports, tracking instances where a human safety driver intervenes to take manual control [18]. Although this framework was designed to enforce transparency, it highlights the systemic challenges of regulatory metric design. Publicly reported data from 2021 reveal extreme variance across developers: within a single reporting period, one entity

recorded 83,423 disengagements across 274 vehicles, whereas another reported only 1,181 disengagements across a significantly larger fleet of 1,978 vehicles [18, 19]. This variation has been attributed substantially to differences in reporting criteria, testing philosophy, and operating environment, although differences in system capability may also contribute [19]. Developers have argued that this metric may not meaningfully distinguish performance between systems.

Additionally, the metric may introduce undesirable behavioral incentives. A low disengagement count is interpreted as evidence of competence. Therefore, the metric creates pressure toward fewer interventions by the safety driver present. A metric that disincentivizes caution in a system where safety is critical is not a neutral instrument.

The reforms subsequently proposed for AV reporting, which include mandatory ODD specification, standardized cause categorization, and quantitative severity grading, show that a severity-blind intervention count is insufficient. Section 4.3 argues that robotic ultrasound should not adopt the unreformed version of this metric.

3.5 Shadow Mode of Autonomous Vehicles

Shadow Mode operation is an economical idea that emerge from AV development. It involves a candidate system that runs alongside the human operator and computes the action it would have taken without actuating anything. Divergences between the human's action and the model's prediction are logged. The vehicles behave as they would without the system installed.

The properties of this operation are attractive for the following reasons. (1) Incremental physical risk is minimal because the system does not act. (2) Data yield is high because operation is continuously instrumented. (3) Collected data concentrate on cases where the model's expectations diverge from expert behavior. Section 6.1 proposes an analogue for ultrasound systems.

3.6 What the Validation Apparatus Supports

In a peer-reviewed analysis covering 56.7 million rider-only miles, Waymo reported statistically significant reduction in crashed vehicle rate relative to human benchmarks across Any-Injury-Reported, Airbag Deployment, and Suspected Serious Injury+ outcomes [20]. Vehicle-to-vehicle intersection crashes represented the largest total reduction, at 96% for Any-Injury-Reported events (87-99%) and 91% for Airbag deployment events (76-98%). Cyclist, motorcycle, pedestrian, secondary crash, and single vehicle crash types were also statistically reduced for the Any-Injury-Reported outcome, and no statistically significant disbenefit was identified in any of the eleven crash type groups examined.

Two features from this validation apparatus must be emphasized. First, tens of millions of miles were required to reach statistical significance for serious outcomes, consistent with the analysis in Section 3.3 and relevant to expectations about the resolving power of small clinical trials. Second, the analysis was conducted within a declared ODD, and its interpretability depends on that declaration.

Section 4: A Proposed Framework for Ultrasound Autonomy

Building on the synthesis in Sections 2 and 3, three conceptual instruments are proposed. Section 4.1 argues that autonomy in ultrasound is not well represented on a single axis and proposed a dual-axis alternative. Section 4.2 defines an ODD specific to ultrasound and identifies a structural property distinguishing it from its automotive counterpart. Section 4.3 proposes a reporting standard intended to avoid the difficulties described in Section 3.4. These are conceptual proposals. None has been empirically evaluated, and Section 7 discusses the limitations this implies.

Section 4.1 The Dual-Axis Autonomy Matrix

Existing autonomy taxonomies, automotive and surgical alike, order systems by how much of a task the machine performs. This ordering is suited to settings in which perception and action can be separated. For AVs, they substantially are the following: the road exists independently of whether the vehicle observes it; a camera that misclassifies a pedestrian has committed a perception error; a vehicle that fails to brake has committed a control error. The two, perception and action, can in principle be diagnosed independently.

Diagnostic ultrasound does not exhibit this separability because the sensor is the actuator. The image acquired is not an observation of a scene existing independently of the observer, but it is rather a physical product of the action that generated it. A poorly chosen probe position does not return a degraded measurement of an available signal. It may return no signal,

through acoustic shadowing behind a rib, dropout through bowel gas, or loss of coupling. The system's perceptual input is generated by its own motor policy, forming a closed loop that neither J3016 nor the surgical taxonomies were constructed to describe [6, 12].

This review, therefore, proposes distinguishing two forms of authority that are largely intertwined in driving but can be assigned independently in ultrasound:

Axis A - Acquisition autonomy (A0 - A5): which agent determines probe placement

1. Axis D - Diagnostic autonomy (D0 - D5): which agent determines the interpretation of the resulting image

Table 1. Proposed levels for the two autonomy axes. Axis A specifies which agent determines probe placement; Axis D specifies which agent determines interpretation of the resulting image. The two axes are treated as independent, so a system occupies a position on each rather than a single autonomy level.

Level	Axis A: Acquisition Autonomy	Axis D: Diagnostic Autonomy
0	Manual Clinician holds and positions the probe throughout.	None System outputs images only. All interpretation by the clinician.
1	Guided System generates positioning cues. Clinician executes all probe motion.	Measurement System computes quantitative measurements, which the clinician verifies.
2	Teleoperated Robot holds and moves the probe under continuous clinician command.	Detection System flags candidate abnormalities for clinician review.
3	Contact-autonomous Robot autonomously regulates contact force and coupling while following a path or target specified by the clinician.	Provisional interpretation System produces a structured finding that the clinician confirms or overrides.
4	Target-autonomous Robot autonomously plans and executes trajectories to reach specified standard views within a defined protocol.	Interpretation with exception review System issues findings, and clinician reviews only flagged cases.
5	Protocol-autonomous Robot determines the examination plan itself, including which views to seek and when the examination is complete.	Autonomous interpretation System issues findings without routine human review.

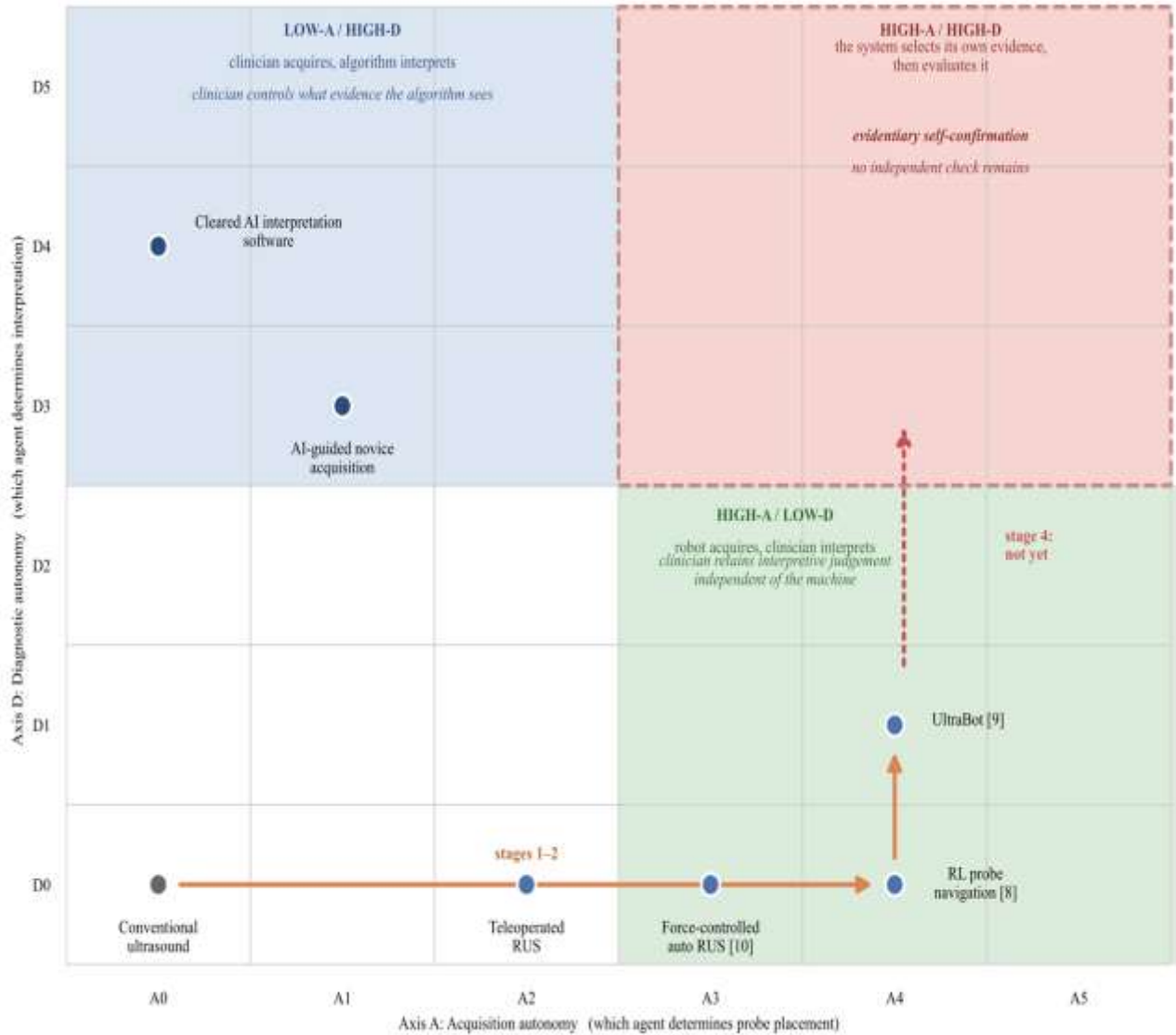
Representative systems positioned on these two axes are shown in Figure 2. The three regions of the matrix carry different risk properties. Systems in the high-A / low-D region, where the robot acquires and the clinician interprets, describe the most current robotic ultrasound research. This configuration retains a clinical judgement that is independent of the machine's interpretation, although not of its acquisition. Systems in the low-A / high-D region, where a clinician acquires and an algorithm interprets, describe most cleared artificial intelligence products in medical imaging. These are safe for the opposite reason, as a human determines is shown. The high-A / high-D configuration, toward which the field is moving, differs from either neighbor in a manner argued here to be qualitative rather than incremental, because such a system selects its own evidence and then evaluates it.

This review proposes the term evidentiary self-confirmation for the associated failure mode. Consider a policy trained end to end against a reward defined on diagnostic outcome. Without additional constraints, such an objective may not distinguish acquiring the view that best reveals the underlying anatomy from acquiring the view most consistent with the conclusion the model is converging upon, since both may increase expected reward. A model exhibiting the latter behavior could display high confidence, high internal consistency across acquired frames, and stable measurements while being incorrect. Its internal agreement would not constitute independent corroboration because the system generated the evidence over which agreement is assessed.

This review argues that evidentiary self-confirmation is a plausible failure mode wherever an acquisition policy is optimized jointly with a diagnostic objective, and that it has limited analogue in systems interpreting images acquired by another agent. No cleared medical imaging system is known to combine these forms of authority in this manner, and no existing autonomy taxonomy contains a category that would identify it. The matrix is proposed in part so that this

configuration is named before it is occupied. The failure mode is argued theoretically here and has not been empirically demonstrated.

Figure 2. The Dual-Axis Autonomy Matrix. Acquisition autonomy (Axis A: which agent determines probe placement) and diagnostic autonomy (Axis D: which agent determines interpretation) are treated as independent, with levels defined in Table 1. Green: the robot acquires and the clinician interprets, so interpretive judgement remains independent of the machine. Blue: the clinician acquires and an algorithm interprets, so the clinician controls what evidence the algorithm is shown. Red: the system both selects and evaluates its own evidence. This configuration is termed “evidentiary self-confirmation” because no system is known to occupy it. Orange arrows show the staged sequence proposed in Section 6.3. The dashed red arrow marks a transition this review argues should not be made.



4.2 Anatomical Operational Design Domain

The anatomical operational design domain (A-ODD) is defined as the specified patient- and environment-level conditions under which a robotic ultrasound system’s performance claim is asserted to hold. Table 2 proposes a minimum classification, with each axis grounded in a physical rather than administrative rationale.

Table 2. Proposed minimum specification for an A-ODD. Each axis is grounded in a physical rather than administrative rationale. Example bounds are illustrative; the specification is intended as a minimum set of dimensions that a performance claim should declare, not as fixed thresholds.

Axis	Example Bound	Physical Rationale
Body habitus	BMI 18-32	Depth-dependent attenuation forces a frequency/penetration tradeoff; deep targets in large habitus may be unreachable at diagnostic resolution
Acoustic window quality	Intercostal spacing; bowel gas burden	Determines whether the target can be imaged at all, independent of policy competence
Surgical and scar history	No prior sternotomy or relevant reconstruction	Scar tissue alters acoustic properties and landmark geometry, potentially invalidating learned spatial priors
Motion state	Breath-hold capable; tremor grade	Visual servoing assumes quasi-static anatomy between control updates
Age band	Adult 18-75	Pediatric anatomy is not a scaled adult; organ proportion and tissue composition differ systematically
Contact tolerance	Pain threshold; skin fragility; anticoagulation	Force-control bounds are population-specific and constitute a safety constraint

The function of an A-ODD parallels that of an ODD because it renders the performance claim falsifiable. Reporting that a carotid system succeeds in 92 percent of cases is difficult to interpret. Reporting that it succeeds in 92 percent of adults aged 18-75 with BMI below 32 and adequate acoustic windows constitutes a claim that can be examined, contested, and expanded. The A-ODD differs from its automotive counterpart in one structural respect, which is argued to justify separate treatment rather than simple renaming.

An AV can determine whether it is inside its ODD by consulting a map. The boundary is external, surveyed in advance, and observable without entering it. A robotic ultrasound system has no direct equivalent. Whether a given patient falls within the system's domain depends on properties, such as acoustic window adequacy, tissue attenuation, gas burden, that are not observable externally and cannot be established without scanning. The domain boundary is discoverable only by crossing it. This is compounded by coupling with no automotive analogue: the out-of-domain condition and the system's inability to detect that condition share a common physical cause. Inadequate acoustic windows do not present as an identifiable signal indicating inadequacy. They present as absence, a region of low echogenicity where a structure is expected, which at the image level is not readily distinguishable from a region correctly imaged and containing no abnormality. This suggests that an unmodified ODD framework is not sufficient for ultrasound and motivates an additional construct.

This review proposes that prior to autonomous acquisition, a system execute a brief, low-force survey sweep whose function is jurisdictional rather than diagnostic. This sweep would estimate acoustic tractability across the intended scan region and return one of three verdicts: in-domain, out-of-domain, or indeterminate, with autonomous acquisition proceeding only under the first. The third verdict is not a failure state but the appropriate output for a patient whose scannability cannot be established, terminating in handoff to a human operator.

The necessary components largely exist. Confidence-based estimates of ultrasound image reliability have been used as a reward signal in reinforcement learning probe navigation, and intercostal path planning work explicitly optimizes trajectories to minimize shadowing [4, 16, 21]. This review proposes that the signals currently used as objectives maximized during scanning would additionally serve as gates evaluated before scanning.

4.3 Avoiding the Disengagement Metric Failure

The ultrasound analogue of the disengagement report is an intervention rate, the number of sonographer takeovers per hundred autonomous examinations. It can be readily collected but would not be comparable across sites with differing supervision cultures. It would disadvantage cautious supervisors and weight a takeover improving image aesthetics identically to one identifying a missed arterial dissection.

This review proposes a three-part reporting standard instead with each element addressing a documented shortcoming of the automotive precedent.

A-ODD stratified performance. Success rates reported per declared domain stratum rather than pooled. A pooled figure represents a weighted average over an unspecified population.

1. Severity graded intervention. A classification based on cosmetic, completeness, measurement, and diagnosis that replaces an undifferentiated count. This parallels the severity grading proposed for AV incident reporting and reduces the incentive to suppress low consequence interventions [19].
2. Out-of-domain refusal rate as a primary reported quantity. For example, a system declining 20 percent of patients and reporting that fact is safer and more clinically useful than one declining none while failing undetectably in 8 percent. Reporting refusal makes appropriate abstention an explicitly measurable component of system performance.

Section 5: Where the Analogy Breaks

This section identifies three aspects in which ultrasound differs structurally from driving. Two render the ultrasound problem more difficult while the third suggests an advantage that is not yet widely explored.

5.1 Limitations of Level 3 Supervision in Ultrasound

Level 3 supervision, also known as conditional automation, is one where the system drives but a human must resume control on request. This means that the system is competent enough that sustained supervision degrades in practice, while the supervisor is required precisely at the moments when the system has encountered a situation it cannot handle.

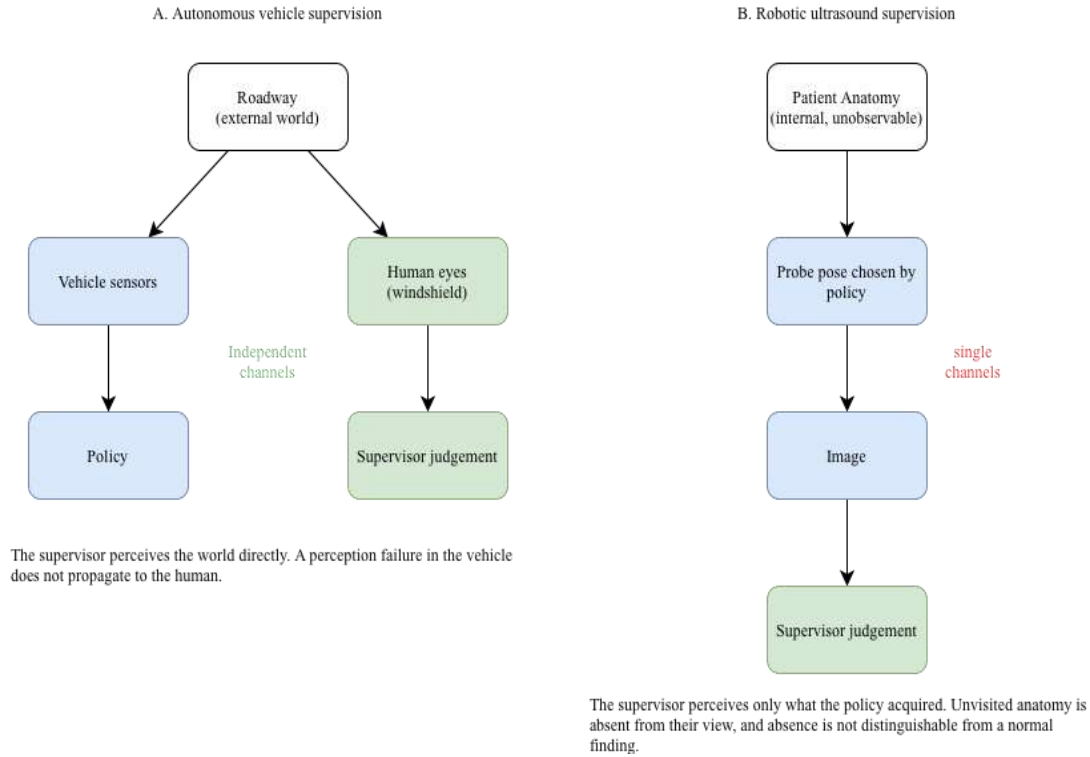
Reported takeover lead times in conditionally automated driving cluster between three and seven seconds [22]. This is a window that must accommodate re-acquisition of situational awareness that has not been continuously maintained. In the investigation of the 2018 Tempe fatality, the NTSB identified the operator's failure to monitor the driving environment as the probable cause, with the developer's inadequate mechanisms for addressing automation complacency among the contributing factors [23].

Robotic ultrasound is developing toward an analogous configuration. Autonomous acquisition must be supervised by a clinician expected to intervene on error. This review argues that ultrasound has an additional structural difficulty.

A safety driver has a windshield. Their perception of the roadway is physically independent of the vehicle's perception of it. A system's failure to classify an obstacle does not propagate to the human observer, who could perceive what the system did not. This independence makes automotive supervision coherent despite its documented failures.

A clinician supervising an autonomous ultrasound has no comparable independent channel. Their perceptual access to the patient's anatomy consists of the image the robot's policy elected to acquire. If the robot's policy searches the wrong intercostal space, the supervisor observes the wrong intercostal space. If the policy fails to probe a specific region, that region is absent from the supervisor's view. Hence, the fallback operator inherits the machine's blind spots and does so independently of their level of attention. Attention, therefore, may not be the binding constraint in this setting.

Figure 3. The no-windshield problem: perceptual independence of the fallback operator. Information flow from the observed environment to the human supervisor in two supervisory configurations. (A) In autonomous driving, the roadway is perceived through two independent channels: the vehicle's sensors feeding its policy, and the supervisor's direct visual access through the windshield. A perception failure in the vehicle does not propagate to the human, who could in principle observe what the system did not. (B) In robotic ultrasound, patient anatomy is internal and cannot be observed directly. The supervisor's only perceptual access is the image produced at the probe pose the policy selected, so the two channels collapse into one.



Three design implications considering these limitations:

1. The interface should instrument the search rather than only the output. If the supervisor cannot independently perceive the anatomy, they should at minimum be able to perceive the policy's behavior through probe pose history, cumulative coverage maps, and explicit indication of unvisited anatomy.
2. Redundancy in ultrasound depends on trajectory, not just repetition. This is because rescanning along the exact same paths reproduces the same acoustic shadows, regardless of the transducer used. To achieve meaningful redundancy, the target region must be scanned from a deliberately different angle. Differences between two distinct angles provide actionable data rather than random errors.
3. Handoffs to supervisors are more effective when planned rather than triggered by an emergency. Handoffs driven by exceptions force a disengaged supervisor to re-engage quickly under time constraints. A structured protocol requiring a clinician to perform a specific, routine confirmation step for every exam keeps them continuously engaged at a predictable time cost.

5.2 The Character of the Long Tail

Unlike AVs, which face an unbounded tail of unpredictable human environments that cannot be fully validated through coverage metrics, ultrasound operates within a much more structured domain. A meta-analysis of 23,882 aortic arches reports the bovine arch variant at 13.6%, and accessory spleens at 14.5%, to less frequent anomalies like a retro aortic left renal vein at 3%. [24, 25, 26].

This contrast must not be overstated. Pathological change, surgical alteration, implanted devices, congenital abnormality, and combinations thereof generate an open-ended component that is not enumerable in the manner normal anatomical variation partially is. Ultrasound may make prevalence-weighted benchmarking substantially more tractable than in autonomous driving. Researchers can sample anatomical variants according to known population base rates, while deliberately oversampling rare conditions and applying statistical reweighting to mirror target demographics.

This connects directly to Section 4.2. Bovine arch prevalence varies significantly by ethnicity, occurring in 26.8% of African populations compared to just 7.5% of Asian populations [24]. Because scenario libraries are not population-neutral, systems validated against them inherit those same demographic biases. Consequently, an A-ODD must explicitly define the

target demographic of its validation cohort. Otherwise, performance claims will implicitly encode the specific population from which the data was sampled.

5.3 Partial Replayability of Acquired Volumes

Unlike autonomous driving, where historical miles cannot be reused to evaluate new policies, ultrasound data offers a partial solution to counterfactual validation. A densely sampled 3D ultrasound volume allows candidate probe poses to be re-rendered offline. Re-rendering cannot perfectly replicate angle-dependent speckle, refraction, or true transducer beam geometry. Despite these limits, policies can be tested against real, unseen patient volumes and scored against expert ground truth from flexible viewpoints. This is a validation strategy that remains largely underutilized.

This review proposes a Retrospective Volume Benchmark. This includes a curated library of annotated, dense 3D ultrasound volumes with A-ODD labels for pre-clinical acquisition policy evaluation. This is especially advantageous because evaluating new policies requires no new patient scans; rare anatomical variants can be oversampled deliberately; and policies can be safely re-tested after model updates. This benchmark is limited by its inability to evaluate force control or patient motion and its reliance on re-rendering at poses far from the original sample path. Though, existing building blocks include interactive ray-traced ultrasound simulation [27]. The main challenge lies in integrating them into a standardized community resource.

Section 6: Implementable Proposals

6.1 Shadow Mode Data Collection

To evaluate autonomous ultrasound acquisition policies safely within clinical workflows, the shadow-mode framework, originally developed for AVs, can be adapted for robotic ultrasound. In this setup, a robotic system is deployed at the patient's bedside to execute a candidate control policy continuously in real time while a human sonographer performs a conventional, manual examination. The robotic end-effector remains completely disengaged and maintains no physical contact with the patient throughout the procedure. At each discrete timestep, the policy receives an observation state consisting of the live B-mode ultrasound image stream captured by the manual probe alongside the probe's six degrees of freedom (6-DoF) spatial pose, recorded concurrently via external optical or electromagnetic tracking systems.

The design carries an important limitation. Because the policy never determines probe motion, it is evaluated only on states the human operator visited, so the procedure measures one-step divergence rather than a closed-loop performance. It does not capture compounding error under autonomous rollout. Shadow mode, therefore, provides a lower bound on the discrepancy between policy and expert behavior, and is best understood as a data collection and screening mechanism rather than a substitute for prospective closed-loop evaluation.

Nonetheless, its advantages remain substantial. Because the candidate policy would not control probe motion or patient contact, such deployment could reduce the incremental physical risk associated with prospective evaluation. Institutional review, patient consent, and data protection requirements would nonetheless continue to apply. Divergences concentrate on anatomies where the policy's expectations fail, addressing the generalization gap outlined in Section 2.2. Given that the UltraBot results suggest demonstration volume is a binding constraint, a mechanism that generates expert demonstrations continuously during examinations already being performed is highly advantageous [5].

6.2 Staged expansion under a predetermined change control plan

The FDA's Predetermined Change Control Plan (PCCP) framework, finalized in December 2024 with implementation guidance issued in August 2025, permits manufacturers to preauthorize specified future model modifications within the original marketing submission [28]. It was developed in part because the prior expectation that algorithms be locked at authorization is difficult to reconcile with systems intended to improve through retraining. Published analyses indicate that PCCPs remain a minority feature among authorized devices using AI, with the majority authorized as locked models [29].

This review argues that robotic ultrasound is a good candidate for this mechanism because the A-ODD supplies the modification protocol with natural structure. Each authorized update could be specified as an A-ODD expansion, such as admitting an additional BMI band, age range, or acoustic window class, with prespecified acceptance criteria evaluated on a Retrospective Volume Benchmark for the newly admitted stratum. This would convert the PCCP from a documentation exercise into a staging plan, providing the regulator with a repeatable evaluation that a locked model submission cannot.

6.3 A staged deployment sequence

Combining the foregoing yields a sequence expressed as a movement through the Dual-Axis Autonomy Matrix. The first stage is a shadow mode data collection, in which the system acts on nothing, and divergence data accumulate across sites. The second is narrow high-A / low-D deployment, which involves autonomous acquisition within a restrictive A-ODD. An example of this is carotid intima-media screening in adults aged 18-75 with confirmed pre scan adequate acoustic windows. The interpretation would be performed by a clinician, along with severity graded intervention reporting and published refusal rate. The third is incremental A-ODD expansion under a PCCP, each expansion prespecified, gated on benchmark performance for the newly admitted stratum, and accompanied by updated stratified reporting. The fourth configuration, high-A / high-D, should not be entered until evidentiary self-confirmation has a validated countermeasure.

Section 7: Limitations

This review is not systematic, and no formal quantitative synthesis was conducted. The Dual-Axis Autonomy Matrix, the A-ODD, pre-scan domain admission, and the Retrospective Volume Benchmark are conceptual proposals that have not been empirically evaluated. Their value is argued rather than demonstrated, and each would require validation before adoption.

CONCLUSION

Robotic ultrasound systems maintain probe contact with high reproducibility and learned acquisition policies are approaching expert-level performance on individual anatomical targets. The field's remaining difficulties appear to lie less in control than in characterizing what has been demonstrated and for which populations.

An adjacent field, autonomous driving, has accumulated considerable relevant experience. What AV development established is largely methodological: that a performance claim without a declared operating domain is not falsifiable, that a mandated metric can reshape the behavior it measures, that direct operational experience cannot practically establish the safety of rare events, and that supervision in which a human is nominally responsible for a system they have ceased to monitor is particularly difficult. This review argues these lessons transfer to robotic ultrasound.

The review argues that the automation ladder does not transfer itself unmodified. The intermediate configuration that automotive engineering found most problematic presents an additional difficulty in ultrasound, because a clinician supervising an autonomous examination has no windshield: their view of the patient is the view the machine selected, and the machine's blind spots become theirs irrespective of attention. Where driving asks how a supervisor's alertness can be maintained, ultrasound must first ask what a supervisor is able to observe.

There is one respect in which ultrasound may hold an advantage it has not yet claimed. Autonomous driving has invested heavily in manufacturing experience it cannot repeat, because a given mile is driven only once. A densely sampled patient volume, appropriately stored, admits partial re-scanning. This can be done through a policy that did not exist when the data were collected, along trajectories not selected at acquisition time.

REFERENCES

1. Evans, K., Roll, S., & Baker, J. (2009). Work-related musculoskeletal disorders (WRMSD) among registered diagnostic medical sonographers and Vascular Technologists. *Journal of Diagnostic Medical Sonography*, 25(6), 287–299. <https://doi.org/10.1177/8756479309351748>
2. Murphey, S. (2017). Work related musculoskeletal disorders in sonography. *Journal of Diagnostic Medical Sonography*, 33(5), 354–369. <https://doi.org/10.1177/8756479317726767>
3. Coffin, C. (2014). Work-related musculoskeletal disorders in sonographers: A review of causes and types of injury and best practices for reducing injury risk. *Reports in Medical Imaging*, 15. <https://doi.org/10.2147/rmi.s34724>
4. Li, K., Wang, J., Xu, Y., Qin, H., Liu, D., Liu, L., & Meng, M. Q.-H. (2021). Autonomous Navigation of an ultrasound probe towards standard scan planes with deep reinforcement learning. *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 8302–8308. <https://doi.org/10.1109/icra48506.2021.9561295>
5. Jiang, H., Zhao, A., Yang, Q., Yan, X., Wang, T., Wang, Y., Jia, N., Wang, J., Wu, G., Yue, Y., Luo, S., Wang, H., Ren, L., Chen, S., Liu, P., Yao, G., Yang, W., Song, S., Li, X., He, K., ... Huang, G. (2025). Towards expert-level autonomous carotid ultrasonography with large-scale learning-based robotic system. *Nature communications*, 16(1), 7893. <https://doi.org/10.1038/s41467-025-62865-w>

6. Yang, G. Z., Cambias, J., Cleary, K., Daimler, E., Drake, J., Dupont, P. E., Hata, N., Kazanzides, P., Martel, S., Patel, R. V., Santos, V. J., & Taylor, R. H. (2017). Medical robotics-Regulatory, ethical, and legal considerations for increasing levels of autonomy. *Science robotics*, 2(4), eaam8638. <https://doi.org/10.1126/scirobotics.aam8638>
7. Attanasio, A., Scaglioni, B., De Momi, E., Fiorini, P., & Valdastrì, P. (2021). Autonomy in Surgical Robotics. *Annual Review of Control, Robotics, and Autonomous Systems*, 4(1), 651–679. <https://doi.org/10.1146/annurev-control-062420-090543>
8. Lee, A., Baker, T. S., Bederson, J. B., & Rapoport, B. I. (2024). Levels of autonomy in FDA-cleared surgical robots: a systematic review. *NPJ digital medicine*, 7(1), 103. <https://doi.org/10.1038/s41746-024-01102-y>
9. American Institute of Ultrasound in Medicine. (2023). AIUM practice principles for work-related musculoskeletal disorder. *Journal of Ultrasound in Medicine*, 42(5), 1139–1157. <https://doi.org/10.1002/jum.16124>
10. Jiang, Z., Salcudean, S. E., & Navab, N. (2023). Robotic ultrasound imaging: State-of-the-art and future perspectives. *Medical image analysis*, 89, 102878. <https://doi.org/10.1016/j.media.2023.102878>
11. Lin, X. X., Li, M. D., Ruan, S. M., Ke, W. P., Zhang, H. R., Huang, H., Wu, S. H., Cheng, M. Q., Tong, W. J., Hu, H. T., He, D. N., Lu, R. F., Lin, Y. D., Kuang, M., Lu, M. D., Chen, L. D., Huang, Q. H., & Wang, W. (2025). Autonomous robotic ultrasound scanning system: a key to enhancing image analysis reproducibility and observer consistency in ultrasound imaging. *Frontiers in robotics and AI*, 12, 1527686. <https://doi.org/10.3389/frobt.2025.1527686>
12. SAE International Recommended Practice, Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles, SAE Standard J3016_202104, Revised April 2021, Issued January 2014, https://doi.org/10.4271/J3016_202104.
13. Huang, D., Yang, C., Zhou, M., Karlas, A., Navab, N., & Jiang, Z. (2025). Robot-assisted deep venous thrombosis ultrasound examination using virtual fixture. *IEEE Transactions on Automation Science and Engineering*, 22, 381–392. <https://doi.org/10.1109/tase.2024.3351076>
14. Von Haxthausen, F., Böttger, S., Wulff, D., Hagenah, J., García-Vázquez, V., & Ipsen, S. (2021). Medical Robotics for Ultrasound Imaging: Current Systems and Future Trends. *Current robotics reports*, 2(1), 55–71. <https://doi.org/10.1007/s43154-020-00037-y>
15. Bi, Y., Jiang, Z., Duellmer, F., Huang, D., & Navab, N. (2024). Machine learning in Robotic Ultrasound Imaging: Challenges and perspectives. *Annual Review of Control, Robotics, and Autonomous Systems*, 7(1), 335–357. <https://doi.org/10.1146/annurev-control-091523-100042>
16. Bi, Y., Qian, C., Zhang, Z., Navab, N., & Jiang, Z. (2026). Autonomous Path planning for intercostal robotic ultrasound imaging using reinforcement learning. *Scientific Reports*, 16(1). <https://doi.org/10.1038/s41598-026-37702-9>
17. Kalra, N., & Paddock, S. M. (2016). Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? *Transportation Research Part A: Policy and Practice*, 94, 182–193. <https://doi.org/10.1016/j.tra.2016.09.010>
18. (2026). (rep.). AUTONOMOUS VEHICLE PERMIT HOLDERS IN CALIFORNIA LOGGED MORE THAN 9 MILLION TEST MILES BETWEEN DECEMBER 1, 2024 AND NOVEMBER 30, 2025. Retrieved from <https://www.dmv.ca.gov/portal/news-and-media/autonomous-vehicle-permit-holders-in-california-logged-more-than-9-million-test-miles-between-december-1-2024-and-november-30-2025/>.
19. Skokan, A., & Mareček, J. (2025a). Could disengagement reports indicate evolution of autonomous vehicles? *Vehicles*, 7(2), 32. <https://doi.org/10.3390/vehicles7020032>
20. Kusano, K. D., Scanlon, J. M., Chen, Y. H., McMurphy, T. L., Gode, T., & Victor, T. (2025). Comparison of Waymo Rider-Only crash rates by crash type to human benchmarks at 56.7 million miles. *Traffic injury prevention*, 26(sup1), S8–S20. <https://doi.org/10.1080/15389588.2025.2499887>
21. Karamalis, A., Wein, W., Klein, T., & Navab, N. (2012). Ultrasound confidence maps using random walks. *Medical image analysis*, 16(6), 1101–1112. <https://doi.org/10.1016/j.media.2012.07.005>
22. Morales-Alvarez, W., Sipele, O., Léberon, R., Tadjine, H. H., & Olaverri-Monreal, C. (2020). Automated driving: A literature review of the take over request in Conditional Automation. *Electronics*, 9(12), 2087. <https://doi.org/10.3390/electronics9122087>
23. National Transportation Safety Board. (2019). *Collision between vehicle controlled by developmental automated driving system and pedestrian, Tempe, Arizona, March 18, 2018* (Highway Accident Report NTSB/HAR-19/03). <https://www.nts.gov/investigations/AccidentReports/Reports/HAR1903.pdf>
24. Popieluszko, P., Henry, B. M., Sanna, B., Hsieh, W. C., Saganiak, K., Pękala, P. A., Walocha, J. A., & Tomaszewski, K. A. (2018). A systematic review and meta-analysis of variations in branching patterns of the adult aortic arch. *Journal of Vascular Surgery*, 68(1). <https://doi.org/10.1016/j.jvs.2017.06.097>
25. Vikse, J., Sanna, B., Henry, B. M., Tattera, D., Sanna, S., Pękala, P. A., Walocha, J. A., & Tomaszewski, K. A. (2017). The prevalence and morphometry of an accessory spleen: A meta-analysis and systematic review of 22,487 patients. *International Journal of Surgery*, 45, 18–28. <https://doi.org/10.1016/j.ijsu.2017.07.045>

26. Hostiuc, S., Rusu, M. C., Negoii, I., Dorobanțu, B., & Grigoriu, M. (2019). Anatomical variants of renal veins: A meta-analysis of prevalence. *Scientific Reports*, 9(1). <https://doi.org/10.1038/s41598-019-47280-8>
27. Mattausch, O., Makhinya, M., & Goksel, O. (2017). Realistic ultrasound simulation of complex surface models using interactive Monte-carlo path tracing. *Computer Graphics Forum*, 37(1), 202–213. <https://doi.org/10.1111/cgf.13260>
28. U.S. Food and Drug Administration. (2024). *Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions: Guidance for industry and Food and Drug Administration staff*. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents>
29. Almarie, B., Gonzalez-Gonzalez, L. F., Dos Santos Barbosa, L. A., Lutz, A., Grosse, U., & Fregni, F. (2025). Machine Learning-Enabled Medical Devices Authorized by the US Food and Drug Administration in 2024: Regulatory Characteristics, Predicate Lineage, and Transparency Reporting. *Biomedicines*, 13(12), 3005. <https://doi.org/10.3390/biomedicines13123005>