algorithms

Review

Sensors, Techniques, and Future Trends of Human-Engagement-Enabled Applications: A Review

Zhuangzhuang Dai [1,*], Vincent Gbouna Zakka [1], Luis J. Manso [1], Martin Rudorfer [1], Ulysses Bernardet [1], Johanna Zumer [2] and Manolya Kavakli-Thorne [3]

[1] Department of Applied AI & Robotics, Engineering and Physical Sciences, Aston University, Birmingham B4 7ET, UK; 220181596@aston.ac.uk (V.G.Z.); l.manso@aston.ac.uk (L.J.M.); m.rudorfer@aston.ac.uk (M.R.); u.bernardet@aston.ac.uk (U.B.)

[2] School of Psychology, Institute of Health and Neurodevelopment, Aston University, Birmingham B4 7ET, UK; j.zumer@aston.ac.uk

[3] Aston Digital Futures Institute, Aston University, Birmingham B4 7ET, UK; m.kavakli-thorne@aston.ac.uk

* Correspondence: z.dai1@aston.ac.uk

Abstract: Human engagement is a vital test research area actively explored in cognitive science and user experience studies. The rise of big data and digital technologies brings new opportunities into this field, especially in autonomous systems and smart applications. This article reviews the latest sensors, current advances of estimation methods, and existing domains of application to guide researchers and practitioners to deploy engagement estimators in various use cases from driver drowsiness detection to human–robot interaction (HRI). Over one hundred references were selected, examined, and contrasted in this review. Specifically, this review focuses on accuracy and practicality of use in different scenarios regarding each sensor modality, as well as current opportunities that greater automatic human engagement estimation could unlock. It is highlighted that multimodal sensor fusion and data-driven methods have shown significant promise in enhancing the accuracy and reliability of engagement estimation. Upon compiling the existing literature, this article addresses future research directions, including the need for developing more efficient algorithms for real-time processing, generalization of data-driven approaches, creating adaptive and responsive systems that better cater to individual needs, and promoting user acceptance.

Keywords: human engagement; sensor-based systems; engagement estimation techniques; literature review

Citation: Dai, Z.; Zakka, V.G.; Manso, L.J.; Rudorfer, M.; Bernardet, U.; Zumer, J.; Kavakli-Thorne, M. Sensors, Techniques, and Future Trends of Human-Engagement-Enabled Applications: A Review. *Algorithms* **2024**, *17*, 560. https://doi.org/10.3390/a17120560

Academic Editors: Grammati Pantziou and Frank Werner

Received: 3 September 2024
Revised: 10 October 2024
Accepted: 10 November 2024
Published: 6 December 2024

1. Introduction

As our reliance on technology has continued to intensify, societies and individuals have become dependent on digital technologies for their health, socialisation, work, and well-being. Re-assessing these reliant relationships by having a closer look at human engagement in this highly fragmented digital culture is important for building a digital future.

Human engagement is a crucial construct in many domains: psychology, sociology, education, cognition, behaviour, and sentiment analysis. Reviews [1–3] have touched upon the conception and quantification of engagement. A universal definition of engagement remains elusive. Nonetheless, all would concur that engagement should be abstracted, scoped, and measured according to its context of use.

The best recognized notion of engagement comes from Sidner et al. [4], "the process by which interactors start, maintain, and end their perceived connections to each other during an interaction." Engagement is an interaction process between two or more parties. The process usually involves at least one human user and other human agents and/or apparatuses, software programs, autonomous agents, interfaces, etc. Albeit a widely used term, engagement is often not clearly defined in the literature among different applications.

This article specifically focuses on human agents in a process of interacting with digital interfaces, apparatuses, software, and autonomous agents in pursuit of a productive goal. For instance, in manual takeover after automated driving [5], a driver may disengage to their current task of text messaging on a smartphone whilst commencing with driving. How to design robots to make them more attractive is also beyond the scope. We emphasize that engagement is a process with many quantifiable sub-processes, such as attention and motor response, as shown in Figure 1. We hypothesize engagement to be both idiothetic and allothetic. That is, the interaction or engagement itself is the driving motivation [6] and the engagement process should aim to achieve a certain goal through extracting information from the environment to fulfil it. Engaging in social media on a phone is positive from a phone app design perspective whilst negative from a workplace productivity perspective. We associate engagement with task and productivity as equivalent to *productive engagement*, proposed by J. Nasir [7]. We examine engagement as sub-processes persisting in humans' perception, appraisal, cognitive, motivation, and motor systems where the feedback cycle time is short. Importantly, this permits measurability and utility of engagement. Therefore, we do not take positive appraisal and cognitive load attention [8] without action or with actions that are not directed towards the interaction partner as engagement.

This review focuses on engagement-enabled applications in which successful modelling and measurement will unlock countless opportunities to enhance intelligent systems, automation, safety, human-centric designs, and well-being in general. A number of applications, such as online learning, driver assistance systems, and human–robot interaction (HRI), are beginning to integrate automatic engagement estimation. They will be further empowered by more accurate, reliable, and privacy-friendly human engagement estimators. Unfortunately, we note a gap in the literature to thoroughly evaluate, contrast, and summarize sensors and techniques for human engagement estimation from an applicability point of view. To this end, this work presents a detailed review of sensing technologies for engagement estimation, common approaches, current practices, and future trends of engagement that inform future applications. The aim is to provide landmarks and guidelines on deploying human engagement estimators for researchers and practitioners in the field. This review is timely in this digital era when sensors, algorithms, and ubiquitous computing have reached a high level of maturity to deliver impact to change and enhance our daily lives.

This article strives to answer the following research questions (RQ):

- RQ1: What sensing modalities are used to measure engagement?
- RQ2: Which application domains are and will benefit from human engagement estimation?
- RQ3: What are the current challenges confronting widespread deployment of engagement estimation systems?

This review presents a retrospective and broad comparative analysis of current techniques and algorithms of human engagement estimation. We focus attention on the fundamental element of this digital world—sensors—and how they measure engagement, their accuracy and limitations are in real-world scenarios, as well as challenges and opportunities of deploying sensor-empowered engagement estimators in various application domains. Our key findings include:

- A comprehensive comparison of commonly used sensors for engagement estimation in terms of accuracy, practicality, and main applications. Results demonstrate disparate sensor requirements among use cases.
- Strengths and weaknesses of existing analytic and data-driven engagement estimation algorithms are reviewed highlighting a trend of exploiting Deep Learning for better accuracy and generalization.
- Our review of existing engagement metrics reveals a gap in the knowledge of how to express engagement to be useful across different applications and to cater for individual needs.

- We predict a prevalence of human engagement estimation in digital engagement, driver assistance, human–robot interaction, and Physiological Feedback and Training; however, reliability, generalization, and user acceptance will remain major issues to address.

The remainder of this paper is structured as follows: Section 2 elucidates the context and scope of the review; Section 3 introduces different sensors and common datasets; Section 4 expands on engagement estimation methods; Section 5 reviews engagement-related metrics; Section 6 gives examples of application domains where human engagement estimation has great impact; Section 7 presents an in-depth discussion on the existing work of human engagement estimation with a focus on identifying accuracy, privacy, and other real-world applicability problems; Section 8 concludes this review by sharing insights in future trends of human-engagement-enabled applications.

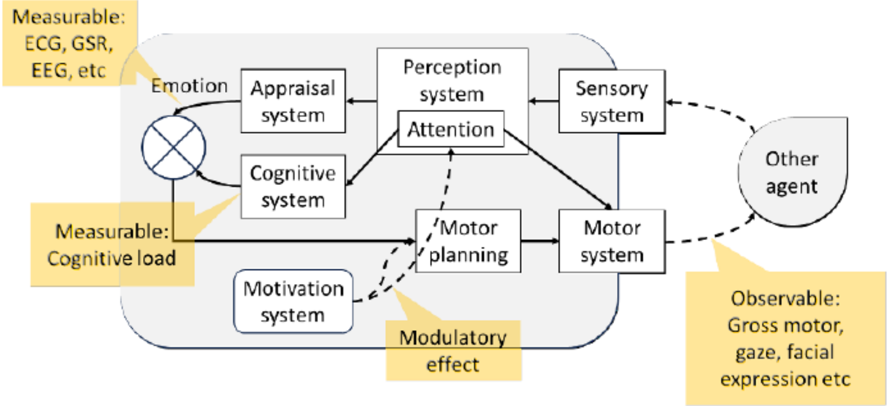


Figure 1. An illustration of the sub-processes of engagement. The human appraisal system, cognitive system, motivation system, and motor system all reveal important information about engagement. Emotions are part of the appraisal process of the behaviour of the interaction partner. A certain level of cognitive load will be associated with engagement. Motor responses such as gross motor, gaze, and facial expression can be captured through observation. Many techniques can measure engagement sub-processes even while not directly measuring engagement as whole.

2. Background

Engagement is often characterised as a variable state which is determined by users' presence, affect, participation, and motion [1]. A popular theory explaining engagement is *flow*, constructed upon the user–system–experience (USE) model [6]. This theory frames engagement as an optimal and enjoyable experience characterised by a tractable challenge, immersion, and immediate feedback [9], which result from users' typology, preferences, and motivations together with system design. Cognitive load theory [8] describes engagement as mental resources of a definite magnitude, which could be drained by poor system design and extraneous distraction. Both theories would back up the variable nature of engagement and its impact on experience, performance, and productivity.

L. J. Corrigan et al. [10] classify engagement into *task engagement* and *social engagement*. Human users' *social engagement*, with counterpart agents for curiosity, companionship, or amusement, is not considered unless networking is the goal per se. One may argue that, e.g., tutors' characters in e-learning or cockpit design for drivers have profound influence on engagement. Hereby, this article concentrates on *task engagement* and *productive engagement* [7] for the purpose of measurability and objectivity of engagement. For instance, we think that engagement in e-learning should be reflected by learning outcomes. Driver engagement comes down to safety, responsiveness, and ability to comply with traffic rules.

Zyngier [11] regards engagement as comprising three perspectives: *behavioural*, *emotional*, and *cognitive*. We follow said taxonomy to identify sensors and to group domains of application.

Behavioural engagement refers to human agents' active participation in the task at hand, such as an operator striking a keyboard, a learner taking notes, and a driver steering the wheel. Audiovisual data provide invaluable information to analyze behavioural factors [12–14], thanks to the rich information contained in the visible and audible spectra. To name a few, eye tracking [12,15], gaze [16,17], posture [18], gesture [4], speech [19–21], and observed behavioural changes [22,23] reveal a large amount of information about users' engagement state.

Emotional engagement describes human agents' affective reactions disclosed by facial expressions in conjunction with body pose and speech [3]. Emotions are part of the appraisal process of the behaviour of the interaction partner. Such emotions can cover all emotion classes/dimensions. Emotion can be part of the idiothetic emotions to appraisal of the engagement process itself. For instance, happiness and contentment usually indicate high-level engagement; boredom or frustration with shaking head often indicate a decrease in engagement. Note that emotion classification is regarded a standalone downstream task often requiring recognition and feature engineering of the Facial Action Unit (FAU). Engagement is correlated to but not necessarily determined by emotion classification. For example, 'surprised' or 'nervous' emotions may indicate strong engagement [24].

Cognitive engagement relates to human agents' internal state of mind. A certain level of cognitive load [8] will be associated with engagement which involves sensory processing and perception. *Cognitive* engagement can be difficult to measure. One can easily relate to the experience of mind-wandering in a class while behaving as if paying attention by gazing at the tutor. Conventionally, questionnaires of self-assessment prevail in harnessing these attributes. Psychophysiological measures [1], such as Electroencephalogram (EEG), Electrocardiogram (ECG), Galvanic Skin Response (GSR), etc., have been commonly used for probing *cognitive* cues in the past decades thanks to the growth of these technologies.

In this review, we differentiate **cues** and **features**, which are often interchangeable in other works. Cue refers to social, physiological, or appearance-based signals. Features stand for digitized attributes of cues to be involved in mathematical computations to derive engagement. We believe distinguishing them helps take the challenge out of understanding which are conceptual and which can be used in computations. Our review concentrates on different sensors' strengths and characteristics in acquiring cues and extracting features. We pay attention to engagement type, among *behavioural*, *emotional*, and *cognitive*, to draw practical implications of deploying sensors in various real-world applications.

3. Sensors

Engagement, as a perceived variable state, can be complicated to measure precisely. Sensors play a critical role in relaying the physiological and cognitive states to a digitized signal to be analyzed. We pay attention to *behavioural*, *emotional*, and *cognitive* cues and features that each type of sensing technique is good at acquiring, as well as strengths and weaknesses of different sensing techniques in different application domains. Specifically, we investigate data logging, eye trackers, cameras for visual surveillance, laser sensors, physiological measurement devices, and microphones.

3.1. Data Logging

Log files, including written records of some attribute as it changes over time and recorded responses to a questionnaire or an interview produced by either participants or observers, are the longest-serving modality for engagement estimation. For good reason, recording data in written format entails the simplest apparatus setup, ranging from interviewing participants and eliciting self-assessment of task engagement to logging behaviours through observers' eyes. Not only simple to conduct, log files support rich description coloured by subjectivity, cognition, emotion, and memory [1].

Many studies to date actively adopt data logging for engagement analysis as they remain efficacious, if not being the only way to unveil engagement. For example, quizzes and assessment performance are typically used to reflect students' level of engagement in e-learning [3]. Nevertheless, logging engagement cannot circumvent the curse of perception bias. That is, participants are often imperfectly poised or motivated to reflect their true engagement state, undermining the authenticity of logged data. Bright learners may perform well in assessment without being engaged in learning. Observers could be deceived by behaviours as no one is able to read minds. Recently, mouse clicks, average time spent on a web page, and other user interaction audits have been exploited to reveal user engagement interacting with digital devices. Obtaining such data is on the rise thanks to a rapid growth in online working, meeting, and learning. However, such data have limitations as they offer rather indirect cues.

3.2. Eye Trackers

Humans rely on their vision in most aspects of their lives. We obtain information about object features (such as size or colour) by looking at them, and we then integrate that information into how we interact with the world. We gather that information continuously by monitoring our environment or, in other terms, simply by looking around. Eye tracking technologies stem from the appeal of understanding human gaze in humans' perception–cognition–action processes. Gaze has long been regarded the most prominent nonverbal communicative signal [25]. Eye tracking helps to reveal a wealth of information to track and quantify human engagement [26].

There are different types of eye trackers, each based on distinct operating principles. In this section, we consider three kinds with a degree of mobility and thus real-world utility: head-mounted eye trackers, screen-based eye trackers, and integrated eye trackers. Head-mounted eye trackers, or eye tracking glasses, are worn by users allowing a range of free movement or unlimited movement [27]. Screen-based eye trackers usually do not have physical contact with users but require calibration to track eye movement on a screen [28]. Note that we exclude eye trackers that infer eye state from ambient cameras through AI. Those will be discussed in the next section. Finally, integrated eye trackers are built into other devices or technologies, such as Virtual Reality (VR) headsets. In clinical or research settings, we can moreover find eye tracking devices that involve attaching a special contact lens or head stabilization equipment, but we do not consider those due to the limitations in deployment.

The eye trackers utilize visible or infrared light [27,29] to infer the direction of human pupils and their projection in the world view, namely gazes and derived fixations. Modern devices offer additional cues, including saccade detection, blink detection, head pose estimation, Region-of-Interest (ROI) event tracking, pupil dilation, and integration of other sensing modalities to fully capture eye state and cues related to human attention. Capturing saccade and microsaccade cues requires a high sampling rate; e.g., high-end eye trackers with over 200 Hz sampling rate such as Tobii Pro series [30] are typically employed. Extra setups are usually needed to acquire head poses, e.g., by deploying QR codes in the environment [31] or using Inertial Measurement Units (IMUs) [32]. Screen-based eye trackers are further constrained in obtaining saccadic and head pose data. Aggregation of eye tracking data may present a rich set of behavioural characteristics, e.g., overlaying fixations to obtain heat maps and scanpath diagrams.

The eye tracking cues yield many features that are directly associated with engagement, as shown in Table 1. Directed gaze refers to gaze landing on an object or group of objects led by an initiator implying immediate intention or interest [33]. Mutual gaze events, i.e., eye contact, have been extensively taken as a sign of bidirectional engagement intention [16,33,34]. R. Kothari et al. [35] investigated coordination of fixation and saccadic events in daily activities. The interplay between fixation and saccades usually reveal a strong indication of a subject's state of concentration as opposed to exploration or mind wandering.

Blink frequency is one of the most far-reaching features in fatigue and drowsiness detection. First proposed in 1990s, PERCLOS [36] measures the percentage over time of eyelid closure over the pupil, reflecting droopy eye closure behaviours. X. Gao et al. [37]

utilized PERCLOS in conjunction with other eye events for driver fatigue monitoring. In the FatigueView dataset [38], PERCLOS realizes reliable performance despite being a generic and computationally inexpensive approach.

Pupil diameter (dilation) links biologically to the arousal system and norepinephrine neurons in the locus coeruleus; pupil diameter is thus a useful independent measure of arousal and engagement that can be obtained from eye trackers beyond the eye position and gaze direction. The pupil diameter links to mind-wandering episodes [39]. Furthermore, the pupil diameter has been shown to be different in individuals with attention-deficit/hyperactivity disorder (ADHD) [40,41], which is another indication that it links with attentional focus.

Table 1. Features that can be extracted from eye-tracking cues and algorithms to derive them.

Cues	Features	Algorithm
Gaze	• 2D/3D Vector • Mutual gaze event (eye contact)	1. Image processing-based pupil tracking [42] 2. Point-of-gaze (POG) estimation from pupil center and corneal reflection [43] 3. Deep Learning methods, e.g., [44]
Fixation	• Total count • *Min*/*max*/*total*/*mean*/*s.d.* (Standard Deviation) of duration • *Min*/*max*/*total*/*mean*/*s.d.* of distance between consecutive fixations	1. Dispersion-duration-based fixation detection [45] 2. Velocity-threshold-based method [46]
Saccade	• Total count • *Min*/*max*/*total*/*mean*/*s.d.* of duration • *Min*/*max*/*total*/*mean*/*s.d.* of saccade amplitude	1. Dispersion-duration-based saccade detection [45] 2. Velocity-threshold-based method [46]
Blink	• Total count • *Min*/*max*/*total*/*mean*/*s.d.* of duration • Frequency (PERCLOS [36])	1. Threshold based pupil confidence method [27] 2. Peak detection in pupil size signal [47] 3. Shape based blink detection [48]
ROI Events	• Total entry/exit count • *Min*/*max*/*total*/*mean*/*s.d.* of dwelling duration	Software-dependent algorithm
Pupil Dilation	• *Min*/*max*/*mean*/*s.d.* of pupil dilation	1. Shape-based pupil boundary detection [49] 2. Edge detection and thresholding [50]
Head Pose [1]	• Pose • *Min*/*max*/*mean*/*s.d.* of angular velocity	1. Pose estimation from visual landmarks [31] 2. Inertial navigation with IMUs [51]

[1] Algorithms are constrained in the scope of head pose detection with the aid of eye trackers.

Nonetheless, all mentioned types of eye trackers are cumbersome in practice. Subjects must wear bespoke glasses or directly face a designated screen. Screen-mounted eye trackers have a very confined effective range. Moreover, calibration per subject per trial is indispensable for most eye-tracking devices on the market. This severely constrains their usability. For instance, safety concerns and inconveniences could prevent usage in driving assistance, outdoor activities, and HRI. Because of these limitations, eye trackers are rarely used outside of controlled laboratory settings [15,28,35,37,52]. The eye movement pattern, modelling, and summaries of data findings are then distilled to formulate design choices; meanwhile, deploying eye trackers in real-world applications remains an expensive and obstructive option. Alternatively, eye state is acquired through other modalities in practice. One of the foremost options is the cameras which will be discussed in the following section.

3.3. Cameras for Visual Surveillance

For the past decades, camera technologies experienced rapid advancement, especially in terms of digitization. Low cost, small form factor, high resolution and frame rate, and reliability all promote the utility of cameras in a great diversity of devices and ubiquitous deployment in manifold applications. Coupled with a surge in computer vision and Machine Learning techniques, cameras have become one of the most widely used sensor modalities for scene understanding and monitoring of human activities.

Traditionally, cameras refer to photo sensors capturing electromagnetic signals in the spectrum visible by humans. Nowadays, technological advances allow cameras to go beyond the visible spectrum. Infrared (IR) cameras capture data in the infrared spectrum, which is just beyond visible [53]. Thermal cameras can detect heat radiation from objects and generate thermal temperature imaging [54], which is helpful to reveal a human's physiological state. RGB-D (Depth) cameras can provide both RGB imaging and depth information [55], allowing the extraction of distance and shape information. Multispectral cameras capture images across multiple wavelengths. Moreover, event-based cameras, also known as neuromorphic cameras, can adapt and respond to brightness changes at pixel level.

Tracking learner engagement with a webcam, a simple RGB camera fitted to a personal computer, produces fruitful results, including large-scale datasets: HBCU [56], DAiSEE [57], EngageWild [58], and EngageNet [59]. Some engagement estimation paradigms rely on extracting facial features, e.g., with OpenFace [60], Dlib [61], etc., to classify emotion prior to deriving engagement level. Popular datasets include AffectNet [62], CK+ [63], and FER2013 [64]. Many studies delve into driver drowsiness and fatigue detection, and also disengagement detection, to facilitate autonomous driving safety assistive systems with in-car cameras [37,65–67]. Datasets like YawDD [68], NTHU-DDD [69], and Fatigue-View [38] are broadly adopted in performance benchmarking. Cameras are also widely adopted in other fields such as child engagement analysis [70,71], ADHD research [72,73], emotion recognition [62–64], social robotics design [34,74], and HRI [13–15]. Details of the aforementioned datasets are summarized in Table 2.

Cameras can capture a wide variety of human engagement cues, as can be seen from Table 3. From Facial Action Units (FAUs) revealing human emotion to body posture and gesture, they can capture almost all appearance-based behaviours if placed at the right position. Taking eye tracking for example, gaze and blinks can be computed given a reasonably good resolution at the cropped eye region in the visual feed [75]. Pin-pointing a fixated object in space is possible with cameras if a 3D mapping of the environment is available [32,76]. Computer-vision-aided head pose estimation has realized superior robustness [60,77]. In fact, many researchers leveraged vision-based methods to label eye state [14,70,78] due to unavailability of eye trackers. Nonverbal communication plays a pivotal role in engagement characterization, and cameras excel at harnessing these personal and sociological cues. Not only can cameras capture body posture, e.g., skeleton-based keypoints [18], but they can detect subtle hand gestures [79], micro-expressions (including involuntary emotional responses) [80], and lip motions which tell who is addressing and controlling the floor. Y. M. Assael et al. [81] proposed a vision-based method to decode speakers' utterances at sentence level. Note that with calibration or depth-sensing ability, a camera can infer distances between agents. Distance is a vital feature correlating to engagement, especially in HRI [82].

Table 2. Vision-based datasets related to engagement estimation.

Dataset	Data Type	Scale	Scenario	Labels
HBCU [56]	Video	25.5 h	E-learning	Engagement degree
DAiSEE [57]	Video	25 h	E-learning	Engagement class
EngageWild [58]	Video	16.5 h	E-learning	Engagement class
EngageNet [59]	Video	31 h	E-learning	Engagement degree
AffectNet [62]	Image	1 M samples	Emotion	Emotion category
CK+ [63]	Video	593	Emotion	Emotion category
FER2013 [64]	Image	35k samples	Emotion	Emotion category
HARMONIC [15]	Multimodal	5 h	HRI	Eye state, EMG, and poses
UE-HRI [78]	Multimodal	54 clips	HRI	Onsets of engagement state
TOGURO [14]	Multimodal	5.83 h	HRI	Engagement level
YawDD [68]	Video	342 clips	Driving	Onsets of drowsiness
NTHU-DDD [69]	Video	450 clips	Driving	Onsets of drowsiness
FatigueView [38]	Video	1384 h	Driving	Eye state, FAU, and onsets

Beyond traditional RGB cameras, applications of RGB-D and thermal cameras are starting to appear. M. Szwoch [83] demonstrates the efficacy of using depth information for facial expression estimation. Recent attempts of using thermal cameras to detect driver disengagement have shown promising results in poorly illuminated conditions [84,85].

Despite the advances and the accuracy of some visual sensing methods, several drawbacks remain. Cameras are known to be prone to performance degradation due to occlusion and sensitivity to illumination conditions [86]. Occluded faces prevent any effective acquisition of FAUs [62]. Glare and darkness can severely affect pose estimation [87], lip reading, and eye state detection. Privacy is another major concern [88], given that humans are not comfortable being monitored with visual surveillance at workplace and at home.

3.4. Laser Sensors

Distance can be a useful factor in human–agent interaction. Especially for humanoid robots, getting close is a first gesture of intent to engage. D. Vaufreydaz et al. [89] used a laser sensor aboard a companion robot to measure distance. The UE-HRI dataset [78] took a distance measure of a laser aboard a Pepper robot as precursor of interaction. Going beyond simple distance measures, Light Detection and Ranging (LiDAR) sensors provide a dense 3D mapping of human agents and their surroundings as point clouds. Researchers have successfully obtained gesture and body poses using LiDAR [90] and even head pose [91]. Nevertheless, LiDAR can at most support engagement estimation, due to its incapacity to detect facial details and emotional cues. Furthermore, powerful LiDARs can be expensive.

3.5. Physiological Measurement Devices

Activity logging, eye tracking, and visual sensing mostly capture observed behaviours. However, certain sub-processes of engagement are internal, and internal state changes and brain activities require physiological sensors attached to human body to measure, such as EEG, ECG, electro-oculogram (EOG), GSR, etc. For instance, the EOG produces Slow Eye Movement (SEM) and Rapid Eye Movement (REM) features [37], which are significant for sleepiness and drowsiness detection but cannot be acquired by mobile eye trackers. A. Chowdhury et al. [2] comprehensively reviewed physiological cues insinuating driver drowsiness, including EEG, ECG, EOG, blinks, SEM, heart rate, blood pressure, respiration rate, Electromyogram (EMG), GSR, and skin temperature. Traditionally, these need to be measured in clinical settings. With the advent of smart wearable devices such as the Fitbit or Apple Watch, some indicators can be continuously monitored throughout daily life. The combination of physiological cues can be even more useful than a single measure in isolation.

3.5.1. Neural EEG Measurements

EEG provides a wealth of neural-based cues that link to attentional engagement, ranging from overt and traditional measures such as overall alpha-band power to more subtle cues/features that have only been extracted recently with modern machine learning.

Cognitively relevant neural oscillations span a range from 4 to 100 Hz and are divided into bands that roughly indicate distinct functions. Theta band (4–8 Hz) power links to memory formation and retrieval [92,93]; a memory is much less likely to be encoded if a person is not engaged. Eye saccades occur in the theta band; however, neural oscillations related to their planning are tied to the alpha band [94], which may then be linked with eye tracking or camera information for eye-gaze tracking. These saccades are often socially relevant [95] and thus a critical feature for social engagement.

The alpha band (8–12 Hz) has been shown to have strong links to control of attention [96], from general engaged/sustained attention [97] to control of spatially allocated attention [98] and selective attention [99]. Clinically relevant, aberrant alpha power is often seen in those with ADHD [100,101].

Beta band (13–30 Hz) power is most linked with motor planning and execution; as such, beta power can predict eye movements or arm reach movements. Even when a movement is only imagined, beta band power strongly links to this imagery/intention and is thus a strong indicator for engagement [102].

The ratio of theta to beta power (theta–beta ratio; TBR) is a clinically relevant metric for identifying those with ADHD [103]. Furthermore, 'engagement' has been quantified as a ratio of beta/(alpha+theta) [104,105] and has been utilized in brain–computer interfaces. Thus, the interplay of the power *across* bands is just as relevant as the power in individual bands for both HRI and clinical purposes.

Finally, the phase (rather than power) of the on-going oscillation has been shown to relate to memory encoding [106] and perception of near-threshold items [97], both of which are critical for engagement with an attention-demanding task such as school learning or safety-flagging (e.g., airport screening).

Furthermore, on-going EEG recordings can provide these relevant cues/features for engagement in real time, unlike in traditional psychology experiments requiring an average of hundreds of repeated trials. At the same time, any of these features or more time-consuming processed derivatives such as connectivity can be directly fed into machine learning algorithms to improve future detection.

The past decade has shown an increase in the development of mobile, portable EEG devices that can provide reliable, robust signals when at home or in natural environments rather than controlled lab settings, such as Neurotechnology, Emotiv, ANT eego, and others, including options for dry connection (i.e., without electrode gel). These together open the gate towards real-world relevant neural measures listed above that link to engagement.

3.5.2. Non-Neural Physiological Measurements

EOG requires two pairs of electrodes applied around the eyes to record both up-down-blink movements and left-right sideways eye movements. An advantage over an eye tracker is that they can give eye position indication even when the eye is nearly closed. However, the eye position from EOG is not as accurate as eye trackers when eyes are open and involve the sensation of electrodes attached to the skin and the wires leading away from the electrodes and thus are only suitable for laboratory setting.

EMG measures muscle activity and can be placed over muscles relevant to the task (e.g., limbs for reaching, driving, etc.; facial muscles for attention and social engagement). Wingenbach [107] discusses benefits and caveats of the use of facial EMG for emotion/affect research. EMG on limbs can be used in the context of simulated driving [108] and driver distraction [109].

GSR has a long history of use in detecting a person's arousal state, in particular related to stress or anxiety. In the context of detecting mind wandering and engagement, GSR is often used alongside other measures. GSR has been shown to be effective for recognizing a distracted state on its own while avoiding the privacy concerns that eye trackers or cameras suffer from [110].

Heart rate, heart rate variability (HRV), and blood pressure can all be measured with a wristwatch-style monitor, also enabling this physiological measure of arousal for engagement purposes (e.g., CogWatch with low-cost research-grade outputs [111]). Specific examples of engagement using cardiac-derived measures include student engagement in different learning education environments [112], task demands interacting with ADHD diagnosis [113], interactions with time-on-task relevant for assessment of prolonged engagement [114], and links with cognition and neurological conditions [115].

One notable study [116] used a combination of neural and non-neural measures to attempt to distinguish between different cognitive states, including workload-focus, distraction, sense of urgency, mind wandering, and cognitive interference. While results were modest, this paves the way for more precise dissociation of cognitive state beyond just on-task or off-task.

However, physiological signals can be intrusive or obstructive in practical applications such as driving [2], although some applications may be possible, e.g., wearable GSR [110] and minimal size slim eye trackers.

3.6. Microphone

Engagement is indeed a multifaceted phenomenon. One may be found with eyes closed and motionless whilst being highly engaged in a phone call with plug-in headphones. Speech and dialogue open up a whole new realm for human engagement estimation. Hands-free phone calls mark a great emancipation for drivers yet engender risks of diverted attention [117]. Whoever leads the floor by actively conversing often exhibits a high level of engagement [118], no matter whether in class or online. Using vocal wake words is, by far, the most popular way to awaken and initiate interactions with smart devices, autonomous agents, and robots [12,19,119]. The microphone is the device to accomplish these tasks. The sound source direction, distance, and multi-speaker separation can all be utilized together in engagement-related design [82]. A microphone array can be leveraged for triangulation [120], i.e., to obtain arriving angle of sound via analyzing the Time-Difference of Arrival (TDOA) of each microphone unit. Distance estimation is also achievable with sound amplitude attenuation modelling [121].

A Speech Activity Detection (SAD) module [89] is often adopted to distinguish speech cues from non-speech acoustic cues. The latter refers to any acoustic cue other than speech itself, such as background noise which reveals speakers' contextual and environmental settings to an extent. Speech cues contain many features: pitch, pitch contours, speech rate, voice intensity, and, most importantly, semantics, which directly relates to speakers' interest, enthusiasm, and immersion in the subject matter [21,25]. It is possible to use even lower-level acoustic features of fundamental frequency, log-energy, and Mel-Frequency Cepstrum Coefficients (MFCCs) for engagement estimation [78] with the aid of sound analysis toolkits like openSMILE [122].

A common paradigm for analyzing multiparty conversational engagement resorts to who leads a conversation [123]. This requires understanding who controls the floor and detecting whenever floor exchanges take place [82]. Considering conversational semantics, an adjacency pair [33], i.e., an utterance following a previous one in response, typically manifests a high engagement level. Backchannel [33] reflects engagement in a similar vein, e.g., a verbal communication of "yeah" or gestural communication of nodding usually resonating to highly engaged interactions.

Nevertheless, speech and audio sensing is regarded among the most privacy-unfriendly means, second to none even including cameras [124]. Speech and non-speech acoustic cues can disclose sensitive attributes [125] about users' personality traits, gender, health condition, geographical origin, socioeconomic status, moods, etc. Users have the right to be informed about the use of microphones and how acquired data are to be stored and processed. Hence, it is compulsory to seek consent before deploying microphones for engagement estimation.

3.7. Summary

The taxonomy of engagement is discussed in [3] following a top-down fashion. A bottom-up approach is adopted in this review to best reflect what sensors can capture which type of *behavioural*, *emotional*, and *cognitive* cues and their indicative application domain, as shown in Table 3. Table 3 also exemplifies accuracy, practicality, and application domains of the sensors reviewed.

Table 3. Sensor modalities useful for human engagement estimation.

	Cues	Type of Engagement	Accuracy †	Practicality ⋆	Application

Data logging	Questionnaire	Cognitive	L	Weak	Learner Assessment;
					User Experience [34];
					Human–Robot Interaction [25]

	Interview	Cognitive	L	Weak	
	Observer Scores	Behavioural	L	Weak	
	Mouse Clicks	Behavioural	M	Strong	
	Web/App Access Duration	Behavioural	M	Strong	

Eye tracker	Gaze	Behavioural	H	Medium	Driving [37];
					Social Interaction [123];
					ADHD Diagnosis [73];
					Human–Robot Interaction [82]

	Fixation	Behavioural	H	Weak	
	Saccade	Behavioural	H	Weak	
	Blink	Behavioural	H	Medium	
	ROI Events	Behavioural	H	Medium	
	Head Pose	Behavioural	M	Strong	

Camera	Facial Action Units	Behavioural and Emotional	H	Strong	Driving [38];
					E-learning [57];
					Social Interaction;
					Human–Robot Interaction [14]

	Sentiment	Emotional	M	Strong	
	Eye State	Behavioural	H	Strong	
	Posture	Behavioural	M	Strong	
	Gesture	Behavioural	M	Strong	
	Head Pose	Behavioural	M	Strong	
	Lip Motion	Behavioural	M	Medium	
	Proximity	Behavioural	L	Strong	
Laser	Proximity	Behavioural	L	Weak	Human–Robot Interaction [91]
	Posture	Behavioural	M	Weak	
	Gesture	Behavioural	M	Weak	
	Head Pose	Behavioural	M	Weak	

Physiological sensor	EEG	Emotional and Cognitive	H	Weak	Driving [37];
					Education [112];
					ADHD Diagnosis [126];
					Social Interactions [113];
					Biofeedback for Training [127]

	EOG	Emotional and Cognitive	H	Weak	
	Facial EMG	Emotional	M	Weak	
	GSR	Emotional and Cognitive	M	Weak	
	Heart Rate	Emotional and Cognitive	M	Medium	
	Blood Pressure	Emotional and Cognitive	M	Medium	
	Respiration	Emotional and Cognitive	L	Medium	
Audio recorder	Speech	Behavioural and Emotional	H	Strong	Social Interaction [21]; Human–Robot Interaction [12]
	Non-speech Acoustic Cue	Behavioural and Emotional	M	Medium	
	Floor	Behavioural	M	Strong	

† Accuracy grade High (H)/Medium (M)/Low (L) is determined by self-reported accuracy performance in the literature [2,3,9,14,25,29,34,37,38,56–58,70,82,105,108,110,128]. Not all references directly deal with engagement estimation. Estimation results of human affects, fatigue, and human activity recognition are accepted as ascribing to sensor cues. In case difficulty of the evaluation/experiment/dataset varies cross-publication, we adjust the accuracy grade depending on a reference's year of release. ⋆ Practicality specifies general friendliness of deployment. We mainly consider five aspects: *form factor, ease to set up, intrusiveness, cost,* and *user acceptability*. As sensor technologies keep evolving, here we take cues extracted from the most advanced model of each sensor to the best of our knowledge [2,21,25,27,30,37,69,78,91,105,108,110].

4. Approaches

This section reviews common approaches to human engagement estimation.

4.1. Model-Driven Approaches

Numerous early works attempted to tackle human engagement with rule-based methods. As a classic rule-based method, PERCLOS denotes the percentage of duration of eye closure. A 50% bar is commonly used to determine driver fatigue [36]. A low blink frequency can be a strong indicator of drowsiness as well. For instance, fewer than ten blinks per minute were regarded as tiredness in the experiments of N. Theresia et al. [129]. Head pose and body posture are commonly used cues. I. Choi and Y. Kim [17] utilized a 30° nodding threshold in a short interval to decide driver drowsiness. They found a 94% accuracy from in-house experiments, which will likely degrade on unseen experiment settings and individuals. M. E. Foster et al. [130] tested and contrasted a rule-based engagement estimator with machine learning models in a robot bartender scenario. Specifically, two rules were applied to distinguish human engagement level: (1) the human head was less than 30 cm from the bar, and (2) the human torso was facing the right side with a rotation angle less than 10°. The authors noted that a rule-based classifier, despite being simple, showed equally competitive performance, a 98% engagement detection rate, in comparison to learning-based methods such as Support Vector Regression (SVR) and Naive Bayes. Nonetheless, all algorithms saw less than 68% accuracy and 0.69 F1-score with novel test data confirming their limitation.

An apparent shortcoming of rule-based engagement estimation lies in a lack of robustness in generalization. The performance of rule-based systems degrades significantly when tested online and in the wild [130]. In multi-party interactions with strong social and spatio-temporal dynamics, the pre-defined structure and prior knowledge of the modelling is easily compromised. It is tough for models to discern how many human agents are potentially engaging and to keep track of them, not to mention predicting their engagement accurately.

Rule-based methods not only fall short in adapting to different tasks but find it difficult to accommodate differences of individual users. Different tasks and environments require distinguishable levels and modes of engagement with varying criticality. For example, drivers are allowed substantial nodding and head turning to look at side mirrors whilst reversing, whereas this may indicate disengagement in soldering electronics. What is more, how humans perceive, receive, and express engagement differs significantly in demographics [1]. Researchers have noticed the effects of many individual factors such as age, gender, body type, proficiency in tasks, etc., on engagement modelling. These are important matters in engagement estimation but easily overlooked in a model-driven approach or essentially too difficult to model.

4.2. Data-Driven Approaches

Data-driven approaches stem from the need for better accuracy and robustness of engagement estimation across different application scenarios. This process is highly coupled with the blossoming of artificial intelligence (AI) in the past decade. AI-based techniques have dramatically boosted applications in various sectors, including business, education, agriculture, transportation, gaming, and robotics. Affective computing is amongst the main beneficiaries. As a result, understanding human *behavioural*, *emotional*, and *cognitive* status from raw sensory data is getting easier and more reliable [3].

Engagement estimation can be formulated as a frame-level classification [131] or regression [14] problem or a time series prediction problem [78] by taking as input a set of features over an elapsed time window, as shown in Figure 2. They can be consolidated with input $X_{t-\tau,t}$, where t is the current time and τ is the observation window length. Engagement is represented as y_t. Thus, engagement estimation is expressed as:

$$y_t = \Theta([x_{t-\tau}, \ldots, x_{t-2}, x_{t-1}, x_t]) \tag{1}$$

where Θ is the learning-based model containing learnable parameters. In a frame-level classification or regression, τ equals 0.

A plethora of research adopted handcrafted features with machine learning algorithms to classify engagement state or evaluate engagement level. These handcrafted features include, but are not limited to, facial keypoints of FAUs, head pose, posture, and other nonverbal signs of communication. There are open-source toolkits to help to extract these features, namely Dlib [132], OpenFace [60], OpenPose [18], etc.

Support Vector Machines (SVMs) [37,82,133] are amongst the most commonly used algorithms. They prove effective in classifying multiparty video conversation [118] realizing a 74% accuracy in six-categorical engagement classification. J. Whitehill et al. [56] investigated SVMs, boosting, and Multinomial Logistic Regression (MLR) for engagement classification. All three methods show over 90% accuracy testing on the HBCU dataset. A. Kaur et al. [58] utilized facial and head pose features of videos and compared SVR, Random Forest, and Long Short-Term Memory (LSTM) for engagement regression. They found Random Forest and LSTM are equally competitive, with 0.09 Mean Squared Error (MSE), superior to 0.15 of SVR. M. E. Foster et al. [130] noted that SVM, Nearest-Neighbour, Decision Trees, and logistic regression delivered indistinguishable performance in their experiments taking as input facial expression, gaze, and posture of human. Hidden Markov Models act as the foundation of many social interaction models underpinning a wide variety of applications [134,135]. This extends to multi-party conversational and behavioural engagement modelling [136,137].

Deep Learning refers to those using Neural Networks (NNs) for representation learning instead of handcrafted features. Deep Learning has the power to learn latent features from raw sensory data such as video, physiological measurement, and sound wave. This gives DNNs an edge over other methods, as a robust representation can be learned in the feature space. This in turn enables transfer learning to exploit pre-trained facial and appearance-based feature extractors for enhanced generalization. J. Hadfield et al. [70] contrasted SVMs and Random Forest to their proposed LSTM model for engagement estimation of child–robot interaction. They found that the LSTM model outperformed the other classic machine learning algorithms. E. Chong et al. [131] presented a bespoke Convolutional Neural Network (CNN) model achieving accurate eye contact detection comparable to human experts. Y. Mitsuzumi and A. Nakazawa [16] backed up this finding in comparing CNN+MLP (Multi-Layer Perceptron) to Generative Adversarial Network (GAN) models. They realized over 80% accuracy on two challenging datasets. L. Huang et al. [138] proposed DRNet, a CNN+MLP architecture, to estimate gaze direction in images, achieving state-of-the-art performance.

The authors of the DAiSEE dataset [57] conducted a comparative study among baseline Deep Learning models. A CNN, specifically InceptionV3 [139], was used to classify engagement at frame level just accomplishing a 47% accuracy of classifying learner engagement state. Engagement could be viewed as a video classification problem by integrating temporal information. The authors tested pre-trained 3D CNNs, specifically C3D [140] and Long-Term Recurrent Convolutional Network (LRCN), raising the accuracy to 56% and 58%, respectively. It is proven that temporal information provides effective engagement estimation and time series modelling is promising. A. Ben-Youssef et al. [13] tested signature models for time series prediction, i.e., Gated Recurrent Unit (GRU) and LSTM. They achieved up to 78% accuracy in challenging multimodal HRI tasks. F. D. Duchetto et al. [14] proposed a CNN+LSTM architecture to regress engagement from a live video feed delivering MSE of 0.126. O. Rudovic et al. [75] leveraged Reinforcement Learning (RL) combined with conventional DNNs to push the boundaries of video-based engagement estimation.

In Figure 2, several popular DNN architectures for engagement estimation are shown. These models vary in complexity, training paradigm, and efficacy in different application domains. The choice should rely on overall consideration of all factors.

In summary, it can be seen from the literature that there is a need for greater accuracy and reliability in engagement estimation approaches. In the review paper [3], the authors remarked on the rise of Deep Learning algorithms, which have been surpassing SVMs and Decision Trees since 2019. This has been verified in this article. Rule-based methods [17,130] become dramatically ineffective on out-of-domain, unseen, and large-scale tests. As data-driven approaches, especially Deep Learning, triumph in tackling complex, real-world, noisy, multimodal engagement estimation tasks, they are setting new heights in datasets (Table 2), albeit scattered in different application domains.

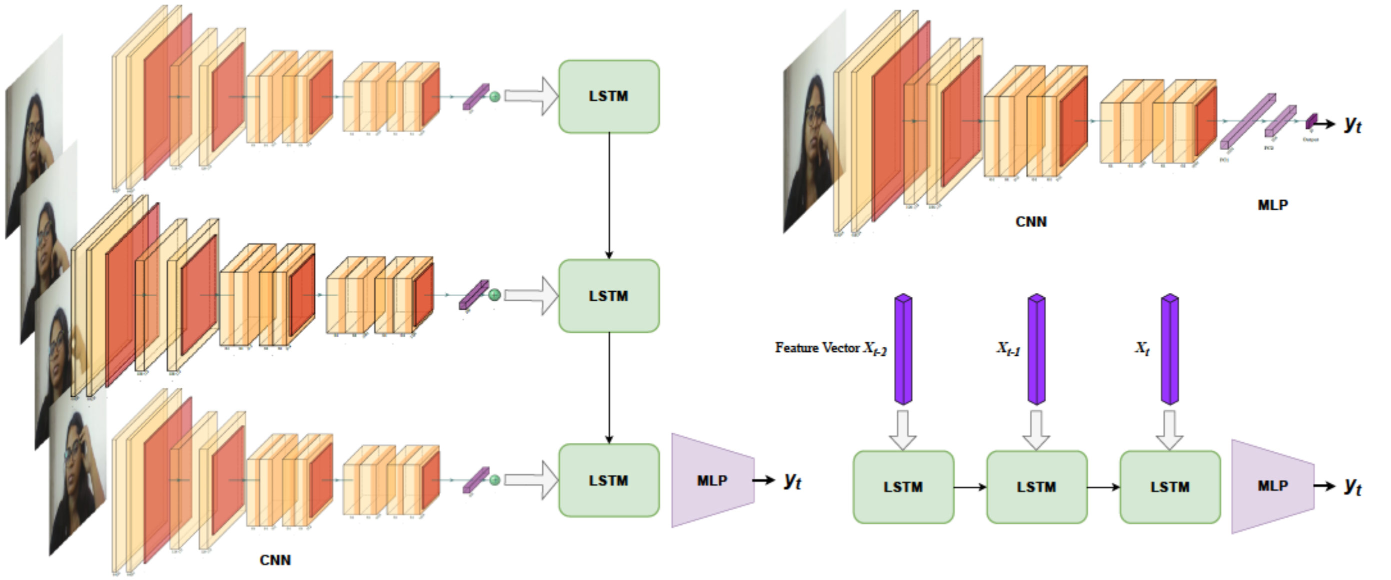


Figure 2. Three typical DNN architectures for human engagement classification or regression. **Left** takes as input a video containing multiple frames and applies CNN+LSTM in conjunction with fully connected layers for engagement output, such as [14,57]. **Top right** evaluates engagement from a single frame with CNN+MLP as is tested in [131,138]. **Bottom right** curates a feature vector concatenating sole or multimodal features before learning hidden states with Recurrent Neural Networks (RNNs), such as GRU and LSTM in [13].

5. Metrics

Metrics play an important role in judging the efficacy of engagement estimators. Comparing estimation results with human performance is a usual means both in performance benchmarking and validation of annotation. Cohen's kappa coefficient, κ, is used in several works [13,56–58] for this purpose.

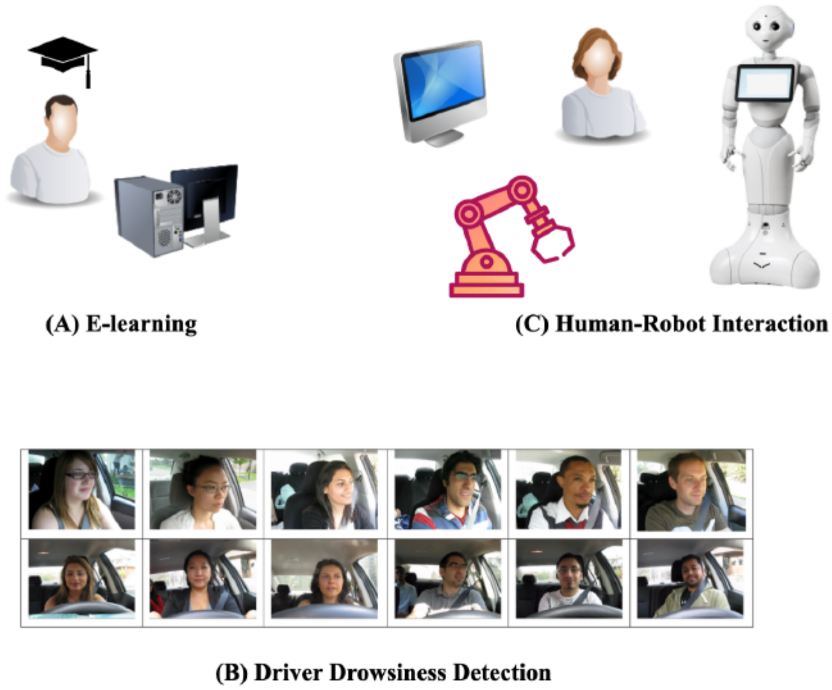
Automatic engagement estimation is regarded as either a classification or a regression task [3] in the vast majority of the literature. Therefore, performance metrics of classification and regression can be used accordingly. For classification, almost all classification-related performance metrics can be applied, including accuracy, precision, recall, F1-score, specificity, sensitivity, Receiver Operating Characteristics (ROCs), Area Under the Curve (AUC), and precision–recall AUC. A confusion matrix can be a useful tool to visualize categorical accuracy and mismatching.

There are different opinions on engagement classes, which can be a major issue in cross-validation and user acceptance. The UE-HRI dataset [78] specifies four engagement labels: *Sign of Engagement Decrease* (SED), *Early sign of future engagement BreakDown* (EBD), *engagement BreakDown* (BD), and *Temporary Disengagement* (TD). J. Whitehill et al. [56] used *Not engaged at all*, *Nominally engaged*, *Engaged in task*, and *Very engaged* in their work. The authors of [58,59] concurred on four levels of engagement: *Not Engaged*, *Barely Engaged*, *Engaged*, and *Highly Engaged*. R. Bednarik et al. [118] defined six shades of engagement, from *No interest* up to *Governing/managing discussion*, in their multi-party conversation study. Four states are used in the DAiSEE dataset [57]: *boredom*, *confusion*, *engagement*, and *frustration*, which mix up engagement with sentiments. This is understandable in e-learning scenarios but does not apply to other applications. These distinct labelling schemes are not only difficult to relate to but barely verifiable in cognitive science. Moreover, they tend to be mutually exclusive and vary from work to work making it harder to adopt, evaluate, and deploy them.

Formulated as a regression task, engagement levels are often estimated as a continuous value between 0 and 1, such as in [14,82]. Metrics used for regression include Mean Absolute Error (MAE), MSE, and Root Mean Squared Error (RMSE), which compute the distance from the predicted value to the ground truth. A major underlying problem of using a floating number is its explainability. It is not clear what it means to be, e.g., 50% engaged. Whether human engagement can be linearly projected is uncertain. Similar to those boundary-blurred classification categories, a value between 0 and 1 lacks cognitive grounding to inform real-world design.

6. Applications

In this section, we review applications driven by human engagement estimation. Since we scope engagement estimation in situations of *humans interacting with digital technologies whilst completing a task*, the counterparts could vary greatly. From smart appliances to touch pads, human engagement estimation could extend to plenty of use cases. Hereby, we focus on four applications as case studies: digital engagement, driver assistance system, human–robot interaction, and Physiological Feedback and Training in ADHD. In these situations, human engagement estimation already starts to realize significance and efficacy. Some examples are shown in Figure 3.



(A) E-learning

(C) Human-Robot Interaction

(B) Driver Drowsiness Detection

Figure 3. Example applications enabled by engagement estimation. (**A**) E-learning can benefit from automatic learner engagement evaluation. (**B**) Driver drowsiness detection is critical for safety and reducing road accidents (data from [68]). (**C**) Engagement estimation plays a key role in designing human–computer interfaces, social robots, and autonomous systems for HRI.

6.1. Digital Engagement

Human engagement is drawing increasing interest due to the surge in digital interfaces, such as personal computers and smart phones, in the modern era. Engagement empowers critical attributes in designing and configuring digital applications ranging from accessibility of a web page to e-learning and online meetings. As specified previously, monitoring learner engagement can be particularly useful for class management and providing live feedback to instructors [56,57]. Using off-the-shelf eye trackers to offer feedback has been shown to be useful in educational settings. For example, students could be prompted to re-focus when attention lapses were detected [141]. Due to the COVID-19 pandemic, there has been a sharp rise in remote working and online meetings which calls for automated participant engagement estimation [142]. R. Garett et al. [143] noted that engagement is an essential requirement, as well as one of the most important benchmarks, in website design. Online interactions for meeting, learning, and entertainment will keep thriving in a digital future. We envision the seamless integration of human engagement estimation in smart devices and software systems will be inevitable to meet this digital culture.

6.2. Driver Assistance System

Distracted and drowsy driving is one of the major factors causing road accidents. In the U.S., it is estimated around 100 thousand crashes resulting in over 1500 deaths and USD 12.5 billion in losses are because of drowsy driving annually [17]. Detection and appropriate intervention in drowsy driving has tremendous societal and economic impact given the huge industry and large number of drivers who can benefit from it worldwide. Consequently, this has been a highly active area of research. Recent research focuses on evaluating driver fatigue and drowsiness in laboratory settings or simulated environments through visual surveillance [65–67,133], eye tracking [37,144], and other physiological measurements [2]. On the other hand, large-scale datasets [85,145] are collected in real-world settings to foster data-driven solutions, such as NTHU-DDD [69] and FatigueView [38]. The Emotion Recognition in the Wild (EmotiW) [146] is a well-recognized event within the research community and includes a driver gaze prediction challenge designated for benchmarking techniques for the online analysis of driver behaviour. It is envisioned more advanced engagement estimators will keep rising in this domain to satisfy the versatile needs for future generations of autonomous vehicles and human–vehicle interfaces.

6.3. Human–Robot Interaction

L. Wang et al. [147] describe various levels of human-robot relationships, from mere coexistence to interaction, cooperation, and finally collaboration. Human engagement can bridge all these relationships [5,148]. In social, industrial, and domestic HRI, identifying and recognizing engaged users acts as a prior for robots to function and to establish situational awareness. In the highest level, *collaboration*, it typically requires a coordinated, synchronous activity to achieve a certain shared goal, and this includes physical contact between human and robot. They highlight the challenges of capturing humans' intents, as well as ensuring safety—human engagement is fundamental for both.

As part of Industry 5.0, it is envisioned that collaborative robots (cobots) will work with humans in close proximity on the shop floor [149]. The inability to perceive engaged users is a long-standing hurdle in designing cobots which can safely contact humans, smoothly switch interlocutor, and collaborate seamlessly with humans to complete tasks. Capturing humans' intents is often performed using gesture recognition [150], implicitly assuming engagement. V. Villani et al. [151] presented a comprehensive review in safety and related standards for industrial cobots. ISO 10218-1/2 [152] is the relevant standard for robots and robotic devices, and ISO 15066 [153] is designed specifically for cobots. Traditionally, safety could be ensured by requiring the human to be outside of the robot's workspace, but this paradigm is broken in human–robot collaboration, where physical contact is desired. Instead, safety is implemented by limiting contact forces [154], safety-rated stops, and speed restriction [149]. For example, ISO 15066 specifies admissible pressures and forces on different areas of the human's body to limit the severeness of possible injuries. We argue that engagement-informed safety mechanisms and protocols are a promising yet under-explored avenue for preventing forceful contact in the first place.

Understanding human attention and intention accurately is paramount for safe and trustworthy HRI in social and domestic context as well. B. A. Newman et al. [15] aimed to use cobots in assisted living with a focus on human intention and mental state modelling. F. Duchetto et al. [14] employed a tour guide robot in a public museum for data collection and user engagement analysis. There has been ongoing research in social and conversational intelligence for years to endow robotic agents to interact with humans in a socially compliant manner [33,78,82,89].

6.4. Physiological Feedback and Training in ADHD

There is wide scope for use of biological-based feedback to help train individuals to boost engagement and become more self-aware of when their attention is about to lapse, not only during measurements but as a training regime that aspires to enable transferable effects beyond monitoring settings.

One specific clinical condition that has had wide testing for biofeedback training is ADHD. A systematic review [127] shows that EEG-based brain–computer interface (BCI) along with a video game can be feasible and effective in improving attention measures in children with ADHD. More broadly, quite a bit of research has examined the possibility of EEG-based neurofeedback as interventions for ADHD, especially targeted at the aforementioned theta–beta ratio (TBR), which has shown promise in some contexts [155], though limitations in others [156,157]. This has potential to wider applications for any work role requiring minimal attention lapses. However, EEG is not always feasible in the home setting.

Potentially more feasible in the non-lab setting is feedback based on eye-tracking. A systematic review of this question [158], which included ADHD but also other neurodevelopmental and neurological conditions, concluded that feedback based on eye-gaze tracking was effective in improving both cognition and emotions.

As discussed in Section 3.2, pupil diameter also links to arousal and engagement state; interesting new findings show that dilation-based biofeedback can enable individuals to control pupil dynamics and linked psychophysiological measures [159]. However, this research is still limited and not yet tested in the context of ADHD or real-world environments. As mentioned above, also relevant for non-lab setting would be a wristwatch-style cardiac measurement that can dissociate engagement with task demands in ADHD [113] or risk of non-engagement as the time on task lengthens [114] and thus could provide biofeedback in these contexts.

7. Discussion

In this article, we present a meta review of latest sensors, current advances of estimation methods, and existing domains of application to guide researchers and practitioners to deploy engagement estimators in various use cases from driver drowsiness detection to HRI. More specifically, this review focuses on accuracy and practicality of use in different scenarios regarding each sensor modality. It is highlighted that multimodal sensor fusion and data-driven methods have shown significant promise in enhancing the accuracy and reliability of engagement estimation.

This section provides a concise retrospect of the literature covered in the previous sections and summarizes our findings in terms of reliability, generalizability, and user acceptance. We particularly focus on the standing challenges of applying engagement estimators to solve real-world problems.

7.1. Reliability

Measuring and detecting human engagement accurately faces multifaceted challenges. A prime concern comes from established datasets. As discussed in Section 5, popular datasets of human engagement adopt naive labels, such as a value between 0 and 1 or binary/four-category/six-category classification of engagement level, without evidencing these labels from a cognitive science perspective. These labels can be overlapping, meaning boundaries between classes, such as *Not engaged* and *Barely engaged*, are fuzzy, which makes a source of confusion for data-driven methods.

Human engagement manifests in different ways, from frequent eyelid closure to head moving away from area of interest. The complexity of human engagement makes it difficult to estimate from one or multiple cues. Taking mind wandering for example, it is difficult to obtain ground truth of the internal state of mind even by interviewing human participants directly or by measuring physiological cues.

There is a rising consensus that the reliable engagement estimation requires multi-modality sensing [12,13,118]. Using multi-modality sensors has the advantage of gaining multidimensional information from the scene allowing collective evaluation and enhanced robustness. For instance, when facial cues are occluded, continued conversation and gesture may be signs of sustained engagement. Thereafter, this complementary information could be merged in the learning-based pipelines through early fusion or late fusion. Early fusion refers to combining low-level sensory features from different modalities at the input level before feeding the concatenated input into a neural network [13,14]. This allows the model to learn joint representations across modalities from raw data with a simple training pipeline. Late fusion processes each modality separately and combines them at the decision level [160,161], e.g., through averaging or voting. However, this means networks for representation learning of each modality must be trained separately. Attention [162] is a common mechanism embedded in state-of-the-art neural networks to focus more on informative modalities in different contexts. There are also hybrid approaches combining elements of both early and late fusion.

Engagement should be carefully put in the context of long-term versus short-term estimation. Diagnosis of ADHD could be a short-term practice, whereas monitoring it is considered a long-term campaign. User adhesiveness to a digital platform is typically deemed long term, but measuring engagement in a single usage is possible. Although attention lapses in class can be short term [12], learning effectiveness and learner's performance are usually assessed in the long run. In contrary, driver drowsiness detection and HRI design demand real-time responsiveness. Drivers should be warned immediately for confirmed disengagement caused by, e.g., sleepiness or fatigue. Social robots are expected to be exceptionally acute in discerning engaged users and floor changes. Real-time engagement recognition is paramount for industrial robotic agents to collaborate with humans safely on the shop floor. There is more to be accomplished to close the gap of real-time engagement estimation.

7.2. Generalization

Engagement is a phenomenon that varies significantly among individuals. Researchers have noted a higher risk of road accidents for those with ADHD [126] and evaluated the impact of ADHD on drivers who obviously exhibit different patterns of engagement and can benefit from bespoke estimators. Similarly, it is found adult learners usually cannot maintain an attention span exceeding 45 min. This varies by age, well-being, and individual circumstances. Having an adaptive and individualized engagement estimator is appealing in many applications. A driver assistance system that prompts felicitously is more acceptable to ADHD drivers. Designing social robots that adapt to individual ways of engagement has always been a prime goal [82]. Allowing calibration and algorithmic adaptability could be the next thing for engagement estimation. Despite successes of data-driven methods, few have accomplished this in current applications.

Engagement is task-oriented. In some situations, having one's eyes glued to a task is critical. Otherwise, this is usually unnecessary. Besides individualization, an emerging focus is the generalization of engagement estimation. The generalization of an engagement estimator is crucial in education, industry, and social settings because it allows for plug and play across diverse tasks and among different users, improving usability of engagement estimation systems and reducing development costs.

Thanks to a fast-growing amount of data in the human engagement research community, learning-based approaches are well poised to thrive. A well-generalized engagement estimator should be able to accommodate versatile tasks, varying attention spans, learning, driving, and working styles, as well as environmental factors that impact the sensing from a noisy manufacturing shop to darkness when driving at night. This ensures that businesses and designers can deploy engagement estimators that are broadly effective and reliable. This is not in conflict with more personalized solutions. Generalization and individualization can be reciprocal because the former leads to technological maturity fostering widespread deployment [82] and the latter promotes user acceptance [163]. Ultimately, individualization will deliver usefulness, satisfaction, and profiting end users with different needs.

7.3. Acceptance

Human engagement estimation systems are rapidly gaining traction in diverse fields such as driver assistance systems, HRI, and digital engagement platforms as discussed in Section 6. However, the successful deployment of these systems hinges not only on their technical accuracy but also on their acceptance by end users. Recent research underscores the importance of incorporating user perspectives early in the design process to overcome adoption barriers. Failure to do so often leads to user resistance and dissatisfaction, impeding the successful deployment of new technologies [164–166]. Therefore, understanding the factors that influence user behaviour, motivation, and decision-making is crucial for fostering acceptance to engagement estimation systems.

The Technology Acceptance Model (TAM) [167], developed by Fred Davis, has been a cornerstone in understanding how users adopt and engage with new technologies. The TAM predicts user acceptance by focusing on key factors like perceived usefulness and perceived ease of use. These factors lead to an individual's attitude towards using the technology, which then influences their behavioural intention to use the system, and ultimately their actual use of it.

In human engagement estimation, integrating the TAM into the design and development process can significantly raise the likelihood of user adoption [168,169]. Perceived usefulness is critical in determining whether users find the technology beneficial in their specific context. For instance, driver assistance systems must allow users to perceive that engagement estimation can enhance their experience by appropriate warnings when detecting driver fatigue or distraction with few false positives [170]. In HRI, robots equipped with engagement estimation capabilities must demonstrate their ability to adapt to human emotions and behaviours [171], thereby improving interaction quality.

Perceived ease of use plays a vital role in user acceptance by minimizing the learning curve associated with new technologies. For example, an engagement estimation system in a digital learning environment that is easy to integrate into existing educational tools [172,173] and requires minimal efforts from educators is more likely to be adopted. Moreover, non-intrusiveness, where the system operates seamlessly without disrupting users' activities, enhances perceived ease of use.

Users' attitudes towards using the system are influenced by factors such as trust, privacy, and ethical considerations [174,175]. In human engagement estimation, these concerns are particularly salient due to the sensitive nature of the data collected. Users may refrain themselves from using the systems because of wariness of their private space being monitoring. Addressing these concerns through transparency, robust data security, and ethical safeguards can positively influence users' attitudes.

Incorporating the TAM into human engagement estimation systems design highlights several challenges that must be addressed to ensure successful adoption [176]. Data privacy and security are paramount [177], given the sensitive nature of the collected data. Any perceived risks of data misuse can significantly hinder acceptance. Additionally, bias and fairness [178] in the algorithms must be carefully managed to prevent skewed or unfair outcomes, which could deter users from adopting the technology. The system's adaptability to different contexts and individual users is also crucial. The varying needs and expectations across different user groups mean that a one-size-fits-all approach is unlikely to succeed. Moreover, regulatory compliance with laws like the GDPR [179] is essential to avoid legal challenges and build trust with users.

The behavioural intention to use an engagement estimation system is shaped by the perceived benefits relative to the effort required. If users believe that the system significantly improves their tasks, they are more likely to use it [163,180]. Thus, how to craft useful and attractive engagement estimators is something crucial for developers and designers to think about.

8. Future Directions

Human engagement is an actively explored research area. Yet, conception of engagement in different contexts, measurement of social/task/multi-party engagement, and realizing impact of engagement in real-world applications have huge room for exploration and improvement. The era of big data and rapid growth of sensor technologies engender bright hopes of applying human engagement estimation for positive changes in the near future.

- K. Doherty and G. Doherty [1] concluded that their work conceptualized engagement, and from then on the focus should be how to measure it and how to realize the value it provides.
- In future research directions, V. Villani et al. [151] predicted a pervasive integration of HRI solutions in industries and emphasized the need for robots to develop cognitive skills to take on more complex tasks.
- J. Duque-Domingo et al. [74] called for constructs of speech and voice recognition, emotion detection, and gaze control systems, that is, a multimodal interaction approach to achieving a better engagement-aware robot design.

Capturing and computationally deriving engagement is not easy. Perceiving human agents' engagement in real-world environments may entail lots of logging and signal processing, sophisticated synthesis of where a person is gazing by reading the eyes, understanding who is initiating a conversation with whom via speech analysis, and being vigilant to other environmental stimuli. The same vision is shared that multi-modality sensing [12,74,151] will play a key role in gathering cues of human behaviour and affect by all means. As a result, the choice of sensors and their pros and cons will be crucial, for which we hope this review provides insights. Also, there is vast space for further innovations to allow data-driven approaches to exploit abundant datasets and develop enhanced paradigms for information fusion.

- J. Nasir et al. [7] share interesting thoughts about future research directions. One is to investigate whether engagement features learned from in situ data, such as image/video/multimodal datasets in specific settings (Table 2) still hold in the progression, i.e., robustness in real time. The paper also describes how to replace human annotation of engagement with automated labelling. This resonates with the initiatives of self-supervised and unsupervised learning in machine learning. The motivation is to exploit the large amount of unlabelled data in the public space to improve training.

Essentially, both ideas revolve around generalization. Generalization is fundamental for many engagement estimation systems to adapt to diverse use cases in various applications. We agree that, just like the rapid evolution of Large Language Models (LLMs), more cross-domain data and superior ways of representation learning are the keys.

Computational cost is often disregarded. As demands rise for deep models of greater accuracy and real-time processing, efficiency will become prominent. Moreover, engagement estimation is expected to become more modularized allowing easy integration into different scenarios [2]. Engagement as a Service (EaaS) will be trending, especially in real-time systems where measuring engagement is critical for safety, well-being, acquiring knowledge, and productivity. This will endow seamless embodiment of engagement enlightening a broad spectrum of applications. In vehicles, personal computers, and robots, EaaS will turn engagement into a beacon of "wake words", making tons of applications not only easy to use but cheaper to run.

- In Karimah and Hasegawa's [3] review, they predict addressing individual differences in personalized engagement will be the next challenge because "engagement cues may take different forms depending on the age range, gender, ethnicity, education" and even "cultures and backgrounds".

- A. Chowdhury et al. [2] remarked physiological sensors were "still not in commercial use due to the invasive nature of biofeedback sensors".
- F. Liu et al. [67] point out that privacy issues are rather under-investigated in current driver fatigue research. They argue scaling down the estimator size to allow edge computing could be a way out in future work.

A personalized, customizable, and intelligent engagement estimation is surely desirable. To accomplish this, a major challenge is how to win user acceptance. The comfort and ease of use are important factors for future wearable sensors. From eye tracking to microphones and from model-driven methods to data-driven methods, how the estimation system is put together should exhibit a certain level of explainability and transparency. No matter whether it is worker disengagement, driver assistance, or ADHD diagnosis, a tuned user engagement profile will become a sensitive piece of personally identifiable information (PII) that demands non-disclosure and protection. In design, users' privacy and trust must be taken as priority, ideally incorporating the principles of TAM.

Author Contributions: Conceptualization, Z.D., U.B. and M.K.-T.; methodology, Z.D., V.G.Z., M.R. and J.Z.; formal analysis, Z.D., L.J.M. and U.B.; investigation, Z.D. and J.Z.; resources, V.G.Z. and J.Z.; data curation, Z.D.; writing—original draft preparation, Z.D., V.G.Z., M.R. and J.Z.; writing—review and editing, L.J.M., U.B. and M.K.-T.; visualization, Z.D. and U.B.; funding acquisition, Z.D. and M.K.-T. All authors have read and agreed to the published version of the manuscript.

Funding: This work is funded by 2022/23 Aston Pump Priming Scheme of Aston University.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Doherty, K.; Doherty, G. Engagement in HCI: Conception, Theory and Measurement. *ACM Comput. Surv.* **2018**, *51*, 1–39. [CrossRef]
2. Chowdhury, A.; Shankaran, R.; Kavakli, M.; Haque, M.M. Sensor Applications and Physiological Features in Drivers' Drowsiness Detection: A Review. *IEEE Sens. J.* **2018**, *18*, 3055–3067. [CrossRef]
3. Karimah, S.; Hasegawa, S. Automatic engagement estimation in smart education/learning settings: A systematic review of engagement definitions, datasets, and methods. *Smart Learn Environ.* **2022**, *9*, 31. [CrossRef]
4. Sidner, C.L.; Lee, C.; Kidd, C.D.; Lesh, N.; Rich, C. Explorations in engagement for humans and robots. *Artif. Intell.* **2005**, *166*, 140–164. [CrossRef]
5. Dogan, E.; Yousfi, E.; Bellet, T.; Tijus, C.; Guillaume, A. Manual takeover after highly automated driving: Influence of budget time and Lane Change Assist on takeover performance. In Proceedings of the 32nd European Conference on Cognitive Ergonomics, ECCE '21, Siena, Italy, 26–29 April 2021; ACM: New York, NY, USA, 2021. [CrossRef]
6. Cowley, B.; Charles, D.; Black, M.; Hickey, R. Toward an understanding of flow in video games. *Comput. Entertain.* **2008**, *6*, 1–27. [CrossRef]
7. Nasir, J.; Bruno, B.; Chetouani, M.; Dillenbourg, P. What if Social Robots Look for Productive Engagement? *Int. J. Soc. Robot.* **2022**, *14*, 55–71. [CrossRef]
8. Oviatt, S. Human-centered design meets cognitive load theory: Designing interfaces that help people think. In Proceedings of the 14th ACM International Conference on Multimedia, MM '06, Santa Barbara, CA, USA, 23–27 October 2006; ACM: New York, NY, USA, 2006; pp. 871–880. [CrossRef]
9. Oertel, C.; Castellano, G.; Chetouani, M.; Nasir, J.; Obaid, M.; Pelachaud, C.; Peters, C. Engagement in Human-Agent Interaction: An Overview. *Front. Robot. AI* **2020**, *7*, 92. [CrossRef]
10. Corrigan, L.; Peters, C.; Castellano, G.; Papadopoulos, F.; Jones, A.; Bhargava, S.; Janarthanam, S.; Hastie, H.; Deshmukh, A.; Aylett, R. Social-Task Engagement: Striking a Balance between the Robot and the Task. In Proceedings of the ICSR 2013 Workshop on Embodied Communication of Goals and Intentions, Bristol, UK, 27–29 October 2013.
11. Zyngier, D. (Re)conceptualising student engagement: Doing education not doing time. *Teach. Teach. Educ. Int. J. Res. Stud.* **2008**, *24*, 1765–1776. [CrossRef]
12. Dai, Z.; Park, J.; Kaszowska, A.; Li, C. Detecting Worker Attention Lapses in Human-Robot Interaction: An Eye Tracking and Multimodal Sensing Study. In Proceedings of the IEEE 28th International Conference on Automation and Computing (ICAC), Birmingham, UK, 30 September–1 October 2023. [CrossRef]
13. Ben Youssef, A.; Varni, G.; Essid, S.; Clavel, C. On-the-Fly Detection of User Engagement Decrease in Spontaneous Human–Robot Interaction Using Recurrent and Deep Neural Networks. *Int. J. Soc. Robot.* **2019**, *11*, 815–828. [CrossRef]

14. Duchetto, F.; Baxter, P.; Hanheide, M. Are You Still With Me? Continuous Engagement Assessment From a Robot's Point of View. *Front. Robot. AI* **2020**, *7*, 116. [CrossRef]
15. Newman, B.A.; Aronson, R.M.; Srinivasa, S.S.; Kitani, K.; Admoni, H. HARMONIC: A multimodal dataset of assistive human–robot collaboration. *Int. J. Robot. Res.* **2021**, *41*, 3–11. [CrossRef]
16. Mitsuzumi, Y.; Nakazawa, A. Eye Contact Detection Algorithms Using Deep Learning and Generative Adversarial Networks. In Proceedings of the 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Miyazaki, Japan, 7–10 October 2018; pp. 3927–3931. [CrossRef]
17. Choi, I.H.; Kim, Y.G. Head pose and gaze direction tracking for detecting a drowsy driver. In Proceedings of the 2014 International Conference on Big Data and Smart Computing (BIGCOMP), Bangkok, Thailand, 15–17 January 2014; pp. 241–244. [CrossRef]
18. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
19. Li, C.; Park, J.; Kim, H.; Chrysostomou, D. How Can I Help You? An Intelligent Virtual Assistant for Industrial Robots. In Proceedings of the Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction, HRI '21, Boulder, CO, USA, 8–11 March 2021; ACM: New York, NY, USA, 2021; pp. 220–224.
20. Céspedes, N.; Hsu, A.; Jones, J.M.; Farkhatdinov, I. A Feasibility Study of a Data-Driven Human-Robot Conversational Interface for Reminiscence Therapy. In Proceedings of the 2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids), Ginowan, Japan, 28–30 November 2022; pp. 708–713. [CrossRef]
21. Bohus, D.; Horvitz, E. Managing Human-Robot Engagement with Forecasts and... Um... Hesitations. In Proceedings of the 16th International Conference on Multimodal Interaction, ICMI '14, Istanbul, Turkey, 12–16 November 2014; ACM: New York, NY, USA, 2014; pp. 2–9. [CrossRef]
22. Hessels, R.S.; Benjamins, J.S.; van Doorn, A.J.; Koenderink, J.J.; Holleman, G.A.; Hooge, I.T.C. Looking behavior and potential human interactions during locomotion. *J. Vis.* **2020**, *20*, 5. [CrossRef] [PubMed]
23. Sullivan, B.; Ludwig, C.J.H.; Damen, D.; Mayol-Cuevas, W.; Gilchrist, I.D. Look-ahead fixations during visuomotor behavior: Evidence from assembling a camping tent. *J. Vis.* **2021**, *21*, 13. [CrossRef] [PubMed]
24. Ekman, P.; Freisen, W.V.; Ancoli, S. Facial signs of emotional experience. *J. Personal. Soc. Psychol.* **1980**, *39*, 1125. [CrossRef]
25. Nakano, Y.I.; Ishii, R. Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In Proceedings of the 15th International Conference on Intelligent User Interfaces, IUI '10, Hong Kong, China, 7–10 February 2010; ACM: New York, NY, USA, 2010; pp. 139–148. [CrossRef]
26. Holmqvist, K.; Andersson, R. *Eye-Tracking: A Comprehensive Guide to Methods, Paradigms and Measures*; Oxford University Press: Oxford, UK, 2011; 560p.
27. Kassner, M.; Patera, W.; Bulling, A. Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-Based Interaction. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, UbiComp '14 Adjunct, Seattle, WA, USA, 13–17 September 2014; ACM: New York, NY, USA, 2014; pp. 1151–1160. [CrossRef]
28. Vaidyanathan, P.; Prud'hommeaux, E.T.; Pelz, J.B.; Alm, C.O. SNAG: Spoken Narratives and Gaze Dataset. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Melbourne, Australia, 15–20 July 2018; pp. 132–137.
29. Hooge, I.T.C.; Niehorster, D.C.; Hessels, R.S.; Benjamins, J.S.; Nyström, M. How robust are wearable eye trackers to slow and fast head and body movements? *Behav. Res. Methods* **2022**, *55*, 4128–4142. [CrossRef]
30. Housholder, A.; Reaban, J.; Peregrino, A.; Votta, G.; Mohd, T.K. Evaluating Accuracy of the Tobii Eye Tracker 5. In Proceedings of the Intelligent Human Computer Interaction: 13th International Conference, IHCI 2021, Kent, OH, USA, 20–22 December 2021; Revised Selected Papers; Springer: Berlin/Heidelberg, Germany, 2021; pp. 379–390. [CrossRef]
31. Wang, J.; Olson, E. AprilTag 2: Efficient and robust fiducial detection. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, Republic of Korea, 9–14 October 2016.
32. Paletta, L.; Santner, K.; Fritz, G.; Mayer, H.; Schrammel, J. 3D attention: Measurement of visual saliency using eye tracking glasses. In Proceedings of the CHI '13 Extended Abstracts on Human Factors in Computing Systems, Paris, France, 27 April–2 May 2013.
33. Rich, C.; Ponsler, B.; Holroyd, A.; Sidner, C.L. Recognizing engagement in human-robot interaction. In Proceedings of the 2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Hong Kong, China, 7–10 February 2010; pp. 375–382. [CrossRef]
34. Kompatsiari, K.; Ciardo, F.; Tikhanoff, V.; Metta, G.; Wykowska, A. It's in the Eyes: The Engaging Role of Eye Contact in HRI. *Int. J. Soc. Robot.* **2021**, *13*, 525–535. [CrossRef]
35. Kothari, R.; Yang, Z.; Kanan, C.; Bailey, R.; Pelz, J.; Diaz, G. Gaze-in-wild: A dataset for studying eye and head coordination in everyday activities. *Sci. Rep.* **2020**, *10*, 2539. [CrossRef]
36. Dinges, D.F.; Grace, R.C. *Perclos: A Valid Psychophysiological Measure of Alertness as Assessed by Psychomotor Vigilance*; Federal Highway Administration Office of Motor Carriers: Washington, DC, USA, 1998.
37. Gao, X.Y.; Zhang, Y.F.; Zheng, W.L.; Lu, B.L. Evaluating driving fatigue detection algorithms using eye tracking glasses. In Proceedings of the 2015 7th International IEEE/EMBS Conference on Neural Engineering (NER), Montpellier, France, 22–24 April 2015; pp. 767–770. [CrossRef]

38. Yang, C.; Yang, Z.; Li, W.; See, J. FatigueView: A Multi-Camera Video Dataset for Vision-Based Drowsiness Detection. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 233–246. [CrossRef]
39. Pelagatti, C.; Binda, P.; Vannucci, M. A closer look at the timecourse of mind wandering: Pupillary responses and behaviour. *PLoS ONE* **2020**, *15*, e0226792. [CrossRef]
40. Shirama, A.; Takeda, T.; Ohta, H.; Iwanami, A.; Toda, S.; Kato, N. Atypical alert state control in adult patients with ADHD: A pupillometry study. *PLoS ONE* **2020**, *15*, e0244662. [CrossRef]
41. Kim, Y.; Kadlaskar, G.; Keehn, R.M.; Keehn, B. Measures of tonic and phasic activity of the locus coeruleus-norepinephrine system in children with autism spectrum disorder: An event-related potential and pupillometry study. *Autism Res. Off. J. Int. Soc. Autism Res.* **2022**, *15*, 2250–2264. [CrossRef] [PubMed]
42. Klaib, A.F.; Alsrehin, N.O.; Melhem, W.Y.; Bashtawi, H.O.; Magableh, A.A. Eye tracking algorithms, techniques, tools, and applications with an emphasis on machine learning and Internet of Things technologies. *Expert Syst. Appl.* **2021**, *166*, 114037. [CrossRef]
43. Guestrin, E.; Eizenman, M. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Trans. Biomed. Eng.* **2006**, *53*, 1124–1133. [CrossRef] [PubMed]
44. Zhang, X.; Sugano, Y.; Fritz, M.; Bulling, A. Appearance-based gaze estimation in the wild. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4511–4520. [CrossRef]
45. Salvucci, D.D.; Goldberg, J.H. Identifying fixations and saccades in eye-tracking protocols. In Proceedings of the Eye Tracking Research & Application, Palm Beach Gardens, FL, USA, 6–8 November 2000.
46. Karpov, A. Automated Classification and Scoring of Smooth Pursuit Eye Movements in Presence of Fixations and Saccades. 2011. Available online: https://api.semanticscholar.org/CorpusID:267915170 (accessed on 1 September 2024).
47. Bergamin, O.; Kardon, R. Latency of the Pupil Light Reflex: Sample Rate, Stimulus Intensity, and Variation in Normal Subjects. *Investig. Ophthalmol. Vis. Sci.* **2003**, *44*, 1546–1554. [CrossRef] [PubMed]
48. Soukupová, T.; Cech, J. Real-Time Eye Blink Detection Using Facial Landmarks. 2016. Available online: https://api.semanticscholar.org/CorpusID:21124316 (accessed on 1 September 2024).
49. Daugman, J. How iris recognition works. *IEEE Trans. Circuits Syst. Video Technol.* **2004**, *14*, 21–30. [CrossRef]
50. Zhu, Z.; Ji, Q. Novel Eye Gaze Tracking Techniques Under Natural Head Movement. *IEEE Trans. Biomed. Eng.* **2007**, *54*, 2246–2260. [CrossRef]
51. Severin, I.C.; Dobrea, D.M. Using Inertial Sensors to Determine Head Motion—A Review. *J. Imaging* **2021**, *7*, 265. [CrossRef]
52. Kumari, N.; Ruf, V.; Mukhametov, S.; Schmidt, A.; Kuhn, J.; Küchemann, S. Mobile Eye-Tracking Data Analysis Using Object Detection via YOLO v4. *Sensors* **2021**, *21*, 7668. [CrossRef]
53. Bañuls, A.; Mandow, A.; Vázquez-Martín, R.; Morales, J.; García-Cerezo, A. Object Detection from Thermal Infrared and Visible Light Cameras in Search and Rescue Scenes. In Proceedings of the 2020 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR), Abu Dhabi, United Arab Emirates, 4–6 November 2020; pp. 380–386. [CrossRef]
54. Saputra, M.R.U.; Trigoni, N.; de Gusmão, P.P.B.; Lu, C.X.; Almalioglu, Y.; Rosa, S.; Chen, C.; Wahlström, J.; Wang, W.; Markham, A. DeepTIO: A Deep Thermal-Inertial Odometry With Visual Hallucination. *IEEE Robot. Autom. Lett.* **2019**, *5*, 1672–1679. [CrossRef]
55. Mur-Artal, R.; Tardós, J.D. ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. *IEEE Trans. Robot.* **2017**, *33*, 1255–1262. [CrossRef]
56. Whitehill, J.; Serpell, Z.; Lin, Y.C.; Foster, A.; Movellan, J.R. The Faces of Engagement: Automatic Recognition of Student Engagementfrom Facial Expressions. *IEEE Trans. Affect. Comput.* **2014**, *5*, 86–98. [CrossRef]
57. Gupta, A.; D'Cunha, A.; Awasthi, K.N.; Balasubramanian, V.N. DAiSEE: Towards User Engagement Recognition in the Wild. *arXiv* **2016**, arXiv:1609.01885.
58. Kaur, A.; Mustafa, A.; Mehta, L.; Dhall, A. Prediction and Localization of Student Engagement in the Wild. In Proceedings of the 2018 Digital Image Computing: Techniques and Applications (DICTA), Canberra, Australia, 10–13 December 2018; pp. 1–8. [CrossRef]
59. Singh, M.; Hoque, X.; Zeng, D.; Wang, Y.; Ikeda, K.; Dhall, A. Do I Have Your Attention: A Large Scale Engagement Prediction Dataset and Baselines. In Proceedings of the 25th International Conference on Multimodal Interaction, ICMI '23, Paris, France, 9–13 October 2023; ACM: New York, NY, USA, 2023; pp. 174–182. [CrossRef]
60. Baltrusaitis, T.; Robinson, P.; Morency, L.P. OpenFace: An open source facial behavior analysis toolkit. In Proceedings of the WACV, IEEE Computer Society, Lake Placid, NY, USA, 7–10 March 2016; pp. 1–10.
61. King, D.E. Dlib-ml: A Machine Learning Toolkit. *J. Mach. Learn. Res.* **2009**, *10*, 1755–1758.
62. Mollahosseini, A.; Hasani, B.; Mahoor, M.H. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Trans. Affect. Comput.* **2017**, *10*, 18–31. [CrossRef]
63. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition—Workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 94–101. [CrossRef]
64. Goodfellow, I.J.; Erhan, D.; Carrier, P.L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.H.; et al. Challenges in Representation Learning: A Report on Three Machine Learning Contests. In *Neural Information Processing*; Lee, M., Hirose, A., Hou, Z.G., Kil, R.M., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 117–124.

65. Murphy-Chutorian, E.; Trivedi, M.M. Head Pose Estimation and Augmented Reality Tracking: An Integrated System and Evaluation for Monitoring Driver Awareness. *IEEE Trans. Intell. Transp. Syst.* **2010**, *11*, 300–311. [CrossRef]
66. Zhao, Z.; Xia, S.; Xu, X.; Zhang, L.; Yan, H.; Xu, Y.; Zhang, Z. Driver Distraction Detection Method Based on Continuous Head Pose Estimation. *Comput. Intell. Neurosci.* **2020**, *2020*, 9606908. [CrossRef]
67. Liu, F.; Chen, D.; Zhou, J.; Xu, F. A Review of Driver Fatigue Detection and Its Advances on the Use of RGB-D Camera and Deep Learning. *Eng. Appl. Artif. Intell.* **2022**, *116*, 105399. [CrossRef]
68. Abtahi, S.; Omidyeganeh, M.; Shirmohammadi, S.; Hariri, B. YawDD: A yawning detection dataset. In Proceedings of the 5th ACM Multimedia Systems Conference, MMSys '14, Singapore, 19–21 March 2014; ACM: New York, NY, USA, 2014; pp. 24–28. [CrossRef]
69. Weng, C.H.; Lai, Y.H.; Lai, S.H. Driver Drowsiness Detection via a Hierarchical Temporal Deep Belief Network. In Proceedings of the ACCV Workshops, Taipei, Taiwan, 20–24 November 2016.
70. Hadfield, J.; Chalvatzaki, G.; Koutras, P.; Khamassi, M.; Tzafestas, C.S.; Maragos, P. A Deep Learning Approach for Multi-View Engagement Estimation of Children in a Child-Robot Joint Attention Task. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 1251–1256.
71. Schodde, T.; Hoffmann, L.; Stange, S.; Kopp, S. Adapt, Explain, Engage—A Study on How Social Robots Can Scaffold Second-language Learning of Children. *J. Hum.-Robot Interact.* **2019**, *9*, 1–27. [CrossRef]
72. Harris, K.R.; Friedlander, B.D.; Saddler, B.; Frizzelle, R.; Graham, S. Self-Monitoring of Attention Versus Self-Monitoring of Academic Performance: Effects Among Students with ADHD in the General Education Classroom. *J. Spec. Educ.* **2005**, *39*, 145–157. [CrossRef]
73. Moro, C.; Nejat, G.; Mihailidis, A. Learning and Personalizing Socially Assistive Robot Behaviors to Aid with Activities of Daily Living. *J. Hum.-Robot Interact.* **2018**, *7*, 1–25. [CrossRef]
74. Duque-Domingo, J.; Gómez-García-Bermejo, J.; Zalama, E. Gaze Control of a Robotic Head for Realistic Interaction With Humans. *Front. Neurorobot.* **2020**, *14*, 34. [CrossRef] [PubMed]
75. Rudovic, O.; Park, H.W.; Busche, J.; Schuller, B.; Breazeal, C.; Picard, R.W. Personalized Estimation of Engagement From Videos Using Active Learning With Deep Reinforcement Learning. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–20 June 2019; pp. 217–226. [CrossRef]
76. Wang, H.; Pi, J.; Qin, T.; Shen, S.; Shi, B.E. SLAM-Based Localization of 3D Gaze Using a Mobile Eye Tracker. In Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications, ETRA '18, Warsaw, Poland, 14–17 June 2018; ACM: New York, NY, USA, 2018. [CrossRef]
77. Ruiz, N.; Chong, E.; Rehg, J.M. Fine-Grained Head Pose Estimation Without Keypoints. In Proceedings of the The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Salt Lake City, UT, USA, 18–22 June 2018.
78. Ben-Youssef, A.; Clavel, C.; Essid, S.; Bilac, M.; Chamoux, M.; Lim, A. UE-HRI: A New Dataset for the Study of User Engagement in Spontaneous Human-robot Interactions. In Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI 2017, Glasgow, UK, 13–17 November 2017; ACM: New York, NY, USA, 2017; pp. 464–472. [CrossRef]
79. De Smedt, Q.; Wannous, H.; Vandeborre, J.P. Skeleton-Based Dynamic Hand Gesture Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1206–1214. [CrossRef]
80. Haggard, E.A.; Isaacs, K.S. Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy. In *Methods of Research in Psychotherapy*; Springer: Boston, MA, USA, 1966; pp. 154–165. [CrossRef]
81. Assael, Y.; Shillingford, B.; Whiteson, S.; de Freitas, N. LipNet: End-to-End Sentence-level Lipreading. *arXiv* **2016**, arXiv:1611.01599.
82. Xu, Q.; Li, L.; Wang, G. Designing engagement-aware agents for multiparty conversations. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '13, Paris, France, 27 April–2 May 2013; ACM: New York, NY, USA, 2013; pp. 2233–2242. [CrossRef]
83. Szwoch, M. On Facial Expressions and Emotions RGB-D Database. In Proceedings of the Beyond Databases, Architectures, and Structures, Ustron, Poland, 27–30 May 2014.
84. Knapik, M.; Cyganek, B. Driver's fatigue recognition based on yawn detection in thermal images. *Neurocomputing* **2019**, *338*, 274–292. [CrossRef]
85. Tashakori, M.; Nahvi, A.; Kiashari, S.E.H. Driver drowsiness detection using facial thermal imaging in a driving simulator. *Proc. Inst. Mech. Eng. Part H J. Eng. Med.* **2022**, *236*, 43–55. [CrossRef] [PubMed]
86. Dai, Z.; Saputra, M.R.U.; Lu, C.X.; Trigoni, N.; Markham, A. Indoor Positioning System in Visually-Degraded Environments with Millimetre-Wave Radar and Inertial Sensors: Demo Abstract. In Proceedings of the SenSys '20, Yokohama, Japan, 16–19 November 2020; ACM: New York, NY, USA, 2020; pp. 623–624. [CrossRef]
87. Xie, Q.; Cheng, T.Y.; Dai, Z.; Tran, V.; Trigoni, N.; Markham, A. Illumination-Aware Hallucination-Based Domain Adaptation for Thermal Pedestrian Detection. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 315–326. [CrossRef]
88. Padilla-López, J.R.; Chaaraoui, A.A.; Flórez-Revuelta, F. Visual privacy protection methods: A survey. *Expert Syst. Appl.* **2015**, *42*, 4177–4195. [CrossRef]
89. Vaufreydaz, D.; Johal, W.; Combe, C. Starting engagement detection towards a companion robot using multimodal features. *Robot. Auton. Syst.* **2016**, *75*, 4–16. [CrossRef]

90. Patil, A.K.; Balasubramanyam, A.; Ryu, J.Y.; B N, P.K.; Chakravarthi, B.; Chai, Y.H. Fusion of Multiple Lidars and Inertial Sensors for the Real-Time Pose Tracking of Human Motion. *Sensors* **2020**, *20*, 5342. [CrossRef]
91. Fürst, M.; Gupta, S.T.P.; Schuster, R.; Wasenmüller, O.; Stricker, D. HPERL: 3D Human Pose Estimation from RGB and LiDAR. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 7321–7327. [CrossRef]
92. Hasselmo, M.E.; Howard Eichenbaum. Hippocampal mechanisms for the context-dependent retrieval of episodes. *Neural Netw.* **2005**, *18*, 1172–1190. [CrossRef]
93. Kerrén, C.; Bree, S.v.; Griffiths, B.J.; Wimber, M. Phase separation of competing memories along the human hippocampal theta rhythm. *eLife* **2022**, 11, e80633. [CrossRef]
94. Pan, Y.; Popov, T.; Frisson, S.; Jensen, O. Saccades are locked to the phase of alpha oscillations during natural reading. *PLoS Biol.* **2023**, *21*, e3001968. [CrossRef] [PubMed]
95. Staudigl, T.; Minxha, J.; Mamelak, A.N.; Gothard, K.M.; Rutishauser, U. Saccade-related neural communication in the human medial temporal lobe is modulated by the social relevance of stimuli. *Sci. Adv.* **2022**, *8*, eabl6037. [CrossRef] [PubMed]
96. Jensen, O.; Mazaheri, A. Shaping functional architecture by oscillatory alpha activity: Gating by inhibition. *Front. Hum. Neurosci.* **2010**, *4*, 186. [CrossRef] [PubMed]
97. Mathewson, K.E.; Gratton, G.; Fabiani, M.; Beck, D.M.; Ro, T. To see or not to see: Prestimulus alpha phase predicts visual awareness. *J. Neurosci. Off. J. Soc. Neurosci.* **2009**, *29*, 2725–2732. [CrossRef] [PubMed]
98. Händel, B.F.; Haarmeier, T.; Jensen, O. Alpha Oscillations Correlate with the Successful Inhibition of Unattended Stimuli. *J. Cogn. Neurosci.* **2011**, *23*, 2494–2502. [CrossRef]
99. Foxe, J.J.; Snyder, A.C. The Role of Alpha-Band Brain Oscillations as a Sensory Suppression Mechanism during Selective Attention. *Front. Psychol.* **2011**, *2*, 154. [CrossRef]
100. Arns, M.; Gunkelman, J.; Breteler, M.; Spronk, D. EEG Phenotypes Predict Treatment Outcome to Stimulants in Children with ADHD. *J. Integr. Neurosci.* **2008**, *07*, 421–438. [CrossRef]
101. Mazaheri, A.; Fassbender, C.; Coffey-Corina, S.; Hartanto, T.A.; Schweitzer, J.B.; Mangun, G.R. Differential Oscillatory Electroencephalogram Between Attention-Deficit/Hyperactivity Disorder Subtypes and Typically Developing Adolescents. *Biol. Psychiatry* **2014**, *76*, 422–429. [CrossRef]
102. Pfurtscheller, G.; Neuper, C. Motor imagery and direct brain-computer communication. *Proc. IEEE* **2001**, *89*, 1123–1134. [CrossRef]
103. Arns, M.; Conners, C.K.; Kraemer, H.C. A decade of EEG Theta/Beta Ratio Research in ADHD: A meta-analysis. *J. Atten. Disord.* **2013**, *17*, 374–383. [CrossRef]
104. Pope, A.T.; Bogart, E.H.; Bartolome, D.S. Biocybernetic system evaluates indices of operator engagement in automated task. *Biol. Psychol.* **1995**, *40*, 187–195. [CrossRef] [PubMed]
105. Coelli, S.; Sclocco, R.; Barbieri, R.; Reni, G.; Zucca, C.; Bianchi, A.M. EEG-based index for engagement level monitoring during sustained attention. In Proceedings of the 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, Italy, 25–29 August 2015; pp. 1512–1515. [CrossRef]
106. Cruzat, J.; Torralba, M.; Ruzzoli, M.; Fernández, A.; Deco, G.; Soto-Faraco, S. The phase of Theta oscillations modulates successful memory formation at encoding. *Neuropsychologia* **2021**, *154*, 107775. [CrossRef] [PubMed]
107. Wingenbach, T.S.H. Facial EMG—Investigating the Interplay of Facial Muscles and Emotions. In *Social and Affective Neuroscience of Everyday Human Interaction: From Theory to Methodology*; Boggio, P.S., Wingenbach, T.S.H., da Silveira Coêlho, M.L., Comfort, W.E., Murrins Marques, L., Alves, M.V.C., Eds.; Springer: Cham, Switzerland, 2023.
108. Balasubramanian, V.; Adalarasu, K. EMG-based analysis of change in muscle activity during simulated driving. *J. Bodyw. Mov. Ther.* **2007**, *11*, 151–158. [CrossRef]
109. Khushaba, R.N.; Kodagoda, S.; Liu, D.; Dissanayake, G. Muscle computer interfaces for driver distraction reduction. *Comput. Methods Programs Biomed.* **2013**, *110*, 137–149. [CrossRef] [PubMed]
110. Dehzangi, O.; Rajendra, V.; Taherisadr, M. Wearable Driver Distraction Identification On-The-Road via Continuous Decomposition of Galvanic Skin Responses. *Sensors* **2018**, *18*, 503. [CrossRef]
111. Dankovich, L.J.; Joyner, J.S.; He, W.; Sesay, A.; Vaughn-Cooke, M. CogWatch: An open-source platform to monitor physiological indicators for cognitive workload and stress. *HardwareX* **2024**, *19*, e00538. [CrossRef]
112. Darnell, D.K.; Krieg, P.A. Student engagement, assessed using heart rate, shows no reset following active learning sessions in lectures. *PLoS ONE* **2019**, *14*, e0225709. [CrossRef]
113. Bellato, A.; Arora, I.; Kochhar, P.; Hollis, C.; Groom, M.J. Indices of Heart Rate Variability and Performance During a Response-Conflict Task Are Differently Associated with ADHD and Autism. *J. Atten. Disord.* **2022**, *26*, 434–446. [CrossRef]
114. Csatho, A.; Van der Linden, D.; Matuz, A. Change in heart rate variability with increasing time-on-task as a marker for mental fatigue: A systematic review. *Biol. Psychol.* **2024**, *185*, 108727. [CrossRef]
115. Arakaki, X.; Arechavala, R.J.; Choy, E.H.; Bautista, J.; Bliss, B.; Molloy, C.; Wu, D.A.; Shimojo, S.; Jiang, Y.; Kleinman, M.T.; et al. The connection between heart rate variability (HRV), neurological health, and cognition: A literature review. *Front. Neurosci.* **2023**, *17*, 1055445. [CrossRef]
116. Scheutz, M.; Aeron, S.; Aygun, A.; De Ruiter, J.; Fantini, S.; Fernandez, C.; Haga, Z.; Nguyen, T.; Lyu, B. Estimating Systemic Cognitive States from a Mixture of Physiological and Brain Signals. *Top. Cogn. Sci.* **2024**, *16*, 485–526. [CrossRef] [PubMed]

117. Iqbal, S.T.; Ju, Y.C.; Horvitz, E. Cars, calls, and cognition: Investigating driving and divided attention. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '10, Atlanta, GA, USA, 10–15 April 2010; ACM: New York, NY, USA, 2010; pp. 1281–1290. [CrossRef]
118. Bednarik, R.; Eivazi, S.; Hradis, M. Gaze and conversational engagement in multiparty video conversation: An annotation scheme and classification of high and low levels of engagement. In Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction, Gaze-In '12, Santa Monica, CA, USA, 26 October 2012; ACM: New York, NY, USA, 2012. [CrossRef]
119. Dutsinma, F.L.I.; Pal, D.; Funilkul, S.; Chan, J.H. A systematic review of voice assistant usability: An ISO 9241–11 approach. *SN Comput. Sci.* **2022**, *3*, 267. [CrossRef] [PubMed]
120. Pavlidi, D.; Griffin, A.; Puigt, M.; Mouchtaris, A. Real-Time Multiple Sound Source Localization and Counting Using a Circular Microphone Array. *IEEE Trans. Audio Speech Lang. Process.* **2013**, *21*, 2193–2206. [CrossRef]
121. Valin, J.M.; Michaud, F.; Rouat, J.; Létourneau, D. Robust sound source localization using a microphone array on a mobile robot. In Proceedings of the 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003) (Cat. No. 03CH37453), Las Vegas, NV, USA, 27 October–1 November 2003; Volume 2, pp. 1228–1233.
122. Eyben, F.; Wöllmer, M.; Schuller, B.W. Opensmile: The munich versatile and fast open-source audio feature extractor. In *ACM Multimedia*; Bimbo, A.D., Chang, S.F., Smeulders, A.W.M., Eds.; ACM: New York, NY, USA, 2010; pp. 1459–1462.
123. Bohus, D.; Horvitz, E. Facilitating multiparty dialog with gaze, gesture, and speech. In Proceedings of the International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction, ICMI-MLMI '10, Beijing, China, 8–12 November 2010; ACM: New York, NY, USA, 2010. [CrossRef]
124. Cheng, P.; Roedig, U. Personal Voice Assistant Security and Privacy—A Survey. *Proc. IEEE* **2022**, *110*, 476–507. [CrossRef]
125. Kröger, J.L.; Lutz, O.H.M.; Raschke, P. Privacy Implications of Voice and Speech Analysis—Information Disclosure by Inference. In *Privacy and Identity Management. Data for Better Living: AI and Privacy: 14th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2.2 International Summer School, Windisch, Switzerland, August 19–23, 2019, Revised Selected Papers*; Springer International Publishing: Cham, Switzerland, 2020; pp. 242–258. [CrossRef]
126. Reimer, B.; Mehler, B.; D'Ambrosio, L.A.; Fried, R. The impact of distractions on young adult drivers with attention deficit hyperactivity disorder (ADHD). *Accid. Anal. Prev.* **2010**, *42*, 842–851. [CrossRef]
127. Cervantes, J.A.; López, S.; Cervantes, S.; Hernández, A.; Duarte, H. Social Robots and Brain–Computer Interface Video Games for Dealing with Attention Deficit Hyperactivity Disorder: A Systematic Review. *Brain Sci.* **2023**, *13*, 1172. [CrossRef]
128. Prajod, P.; Lavit Nicora, M.; Malosio, M.; André, E. Gaze-based Attention Recognition for Human-Robot Collaboration. In Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments, PETRA '23, Corfu, Greece, 5–7 July 2023; ACM: New York, NY, USA, 2023; pp. 140–147. [CrossRef]
129. Pasaribu, N.T.B.; Prijono, A.; Ratnadewi, R.; Adhie, R.P.; Felix, J. Drowsiness Detection According to the Number of Blinking Eyes Specified From Eye Aspect Ratio Value Modification. In *Advances in Social Science, Education and Humanities Research; Proceedings of the 1st International Conference on Life, Innovation, Change and Knowledge (ICLICK 2018)*; Atlantis Press: Amsterdam, The Netherlands, 2019; pp. 171–174. [CrossRef]
130. Foster, M.E.; Gaschler, A.; Giuliani, M. Automatically Classifying User Engagement for Dynamic Multi-party Human–Robot Interaction. *Int. J. Soc. Robot.* **2017**, *9*, 659–674. [CrossRef]
131. Chong, E.; Clark-Whitney, E.; Southerland, A.; Stubbs, E.; Miller, C.; Ajodan, E.L.; Silverman, M.R.; Lord, C.; Rozga, A.; Jones, R.M.; et al. Detection of eye contact with deep neural networks is as accurate as human experts. *Nat. Commun.* **2020**, *11*, 6386. [CrossRef]
132. Hasnine, M.; Nguyen Tan, H.; Tran, T.; Bui, T.; Akçapınar, G.; Ueda, H. A Real-Time Learning Analytics Dashboard for Automatic Detection of Online Learners' Affective States. *Sensors* **2023**, *23*, 4243. [CrossRef]
133. Pauly, L.; Sankar, D. Detection of drowsiness based on HOG features and SVM classifiers. In Proceedings of the 2015 IEEE International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), Kolkata, India, 20–22 November 2015; pp. 181–186. [CrossRef]
134. Brand, M.; Oliver, N.; Pentland, A. Coupled hidden Markov models for complex action recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico, 17–19 June 1997; pp. 994–999. [CrossRef]
135. Figueroa-Angulo, J.I.; Savage, J.; Bribiesca, E.; Escalante, B.; Sucar, L.E. Compound Hidden Markov Model for Activity Labelling. *Int. J. Intell. Syst.* **2015**, *05*, 177–195. [CrossRef]
136. Bohus, D.; Horvitz, E. Models for multiparty engagement in open-world dialog. In Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL '09, London, UK, 11–12 September 2009; pp. 225–234.
137. Mihoub, A.; Bailly, G.; Wolf, C. Social Behavior Modeling Based on Incremental Discrete Hidden Markov Models. In *Human Behavior Understanding*; Salah, A.A., Hung, H., Aran, O., Gunes, H., Eds.; Springer: Cham, Switzerland, 2013; pp. 172–183.
138. Huang, L.; Li, Y.; Wang, X.; Wang, H.; Bouridane, A.; Chaddad, A. Gaze Estimation Approach Using Deep Differential Residual Network. *Sensors* **2022**, *22*, 5462. [CrossRef] [PubMed]
139. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 2818–2826.

140. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning Spatiotemporal Features with 3D Convolutional Networks. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 4489–4497. [CrossRef]
141. D'Mello, S.K. Gaze-Based Attention-Aware Cyberlearning Technologies. In *Mind, Brain and Technology: Learning in the Age of Emerging Technologies*; Parsons, T.D., Lin, L., Cockerham, D., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 87–105. [CrossRef]
142. Watanabe, K.; Tanuja Sathyanarayana, T.S.; Dengel, A.; Ishimaru, S. EnGauge: Engagement Gauge of Meeting Participants Estimated by Facial Expression and Deep Neural Network. *IEEE Access* **2023**, *11*, 52886–52898. [CrossRef]
143. Garett, R.; Chiu, J.; Zhang, L.; Young, S.D. A Literature Review: Website Design and User Engagement. *Online J. Commun. Media Technol.* **2016**, *63*, 1–14. [CrossRef]
144. Friedrichs, F.; Yang, B. Camera-based drowsiness reference for driver state classification under real driving conditions. In Proceedings of the 2010 IEEE Intelligent Vehicles Symposium, La Jolla, CA, USA, 21–24 June 2010; pp. 101–106. [CrossRef]
145. Ghoddoosian, R.; Galib, M.; Athitsos, V. A Realistic Dataset and Baseline Temporal Model for Early Drowsiness Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Long Beach, CA, USA, 16–20 June 2019.
146. Dhall, A.; Sharma, G.; Goecke, R.; Gedeon, T. EmotiW 2020: Driver Gaze, Group Emotion, Student Engagement and Physiological Signal based Challenges. In Proceedings of the 2020 International Conference on Multimodal Interaction, ICMI '20, Utrecht, The Netherlands, 25–29 October 2020; ACM: New York, NY, USA, 2020; pp. 784–789. [CrossRef]
147. Wang, L.; Gao, R.; Váncza, J.; Krüger, J.; Wang, X.V.; Makris, S.; Chryssolouris, G. Symbiotic human-robot collaborative assembly. *CIRP Ann.* **2019**, *68*, 701–726. [CrossRef]
148. Wang, X.; Guo, S.; Xu, Z.; Zhang, Z.; Sun, Z.; Xu, Y. A Robotic Teleoperation System Enhanced by Augmented Reality for Natural Human–Robot Interaction. *Cyborg Bionic Syst.* **2024**, *5*, 0098. [CrossRef]
149. Park, J.; Carøe Sørensen, L.; Faarvang Mathiesen, S.; Schlette, C. A Digital Twin-based Workspace Monitoring System for Safe Human-Robot Collaboration. In Proceedings of the 2022 10th International Conference on Control, Mechatronics and Automation (ICCMA), Luxembourg, 9–12 November 2022; pp. 24–30. [CrossRef]
150. Liu, H.; Wang, L. Gesture recognition for human-robot collaboration: A review. *Int. J. Ind. Ergon.* **2018**, *68*, 355–367. [CrossRef]
151. Villani, V.; Pini, F.; Leali, F.; Secchi, C. Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications. *Mechatronics* **2018**, *55*, 248–266. [CrossRef]
152. International Organization for Standardization. *ISO 10218-1:2011 Robots and Robotic Devices—Safety Requirements for Industrial Robots—Part 1: Robots; ISO 10218-2:2011 Robots and Robotic Devices—Safety Requirements for Industrial Robots—Part 2: Robot Systems and Integration*; International Standard: Geneva, Switzerland, 2011.
153. International Organization for Standardization. *ISO/TS 15066:2016 Robots and Robotic Devices—Collaborative Robots*; Technical Specification: Geneva, Switzerland, 2016.
154. Mohammed, A.; Schmidt, B.; Wang, L. Active Collision Avoidance for Human–robot Collaboration Driven by Vision Sensors. *Int. J. Comput. Integr. Manuf.* **2017**, *30*, 970–980. [CrossRef]
155. Gevensleben, H.; Holl, B.; Albrecht, B.; Vogel, C.; Schlamp, D.; Kratz, O.; Studer, P.; Rothenberger, A.; Moll, G.H.; Heinrich, H. Is neurofeedback an efficacious treatment for ADHD? A randomised controlled clinical trial. *J. Child Psychol. Psychiatry Allied Discip.* **2009**, *50*, 780–789. [CrossRef]
156. Cortese, S.; Ferrin, M.; Brandeis, D.; Holtmann, M.; Aggensteiner, P.; Daley, D.; Santosh, P.; Simonoff, E.; Stevenson, J.; Stringaris, A.; et al. Neurofeedback for Attention-Deficit/Hyperactivity Disorder: Meta-Analysis of Clinical and Neuropsychological Outcomes From Randomized Controlled Trials. *J. Am. Acad. Child Adolesc. Psychiatry* **2016**, *55*, 444–455. [CrossRef]
157. Bazanova, O.M.; Auer, T.; Sapina, E.A. On the Efficiency of Individualized Theta/Beta Ratio Neurofeedback Combined with Forehead EMG Training in ADHD Children. *Front. Hum. Neurosci.* **2018**, *12*, 3. [CrossRef] [PubMed]
158. Carelli, L.; Solca, F.; Tagini, S.; Torre, S.; Verde, F.; Ticozzi, N.; Ferrucci, R.; Pravettoni, G.; Aiello, E.N.; Silani, V.; et al. Gaze-Contingent Eye-Tracking Training in Brain Disorders: A Systematic Review. *Brain Sci.* **2022**, *12*, 931. [CrossRef] [PubMed]
159. Meissner, S.N.; Bächinger, M.; Kikkert, S.; Imhof, J.; Missura, S.; Carro Dominguez, M.; Wenderoth, N. Self-regulating arousal via pupil-based biofeedback. *Nat. Hum. Behav.* **2024**, *8*, 43–62. [CrossRef] [PubMed]
160. Dai, Z.; Tran, V.; Markham, A.; Trigoni, N.; Rahman, M.A.; Wijayasingha, L.N.S.; Stankovic, J.; Li, C. EgoCap and EgoFormer: First-Person Image Captioning with Context Fusion. 2022. Available online: https://ssrn.com/abstract=4259901 (accessed on 1 September 2024).
161. Ni, J.; Bai, Y.; Zhang, W.; Yao, T.; Mei, T. Deep Equilibrium Multimodal Fusion. *arXiv* **2023**, arXiv:2306.16645.
162. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.U.; Polosukhin, I. Attention is All you Need. In *Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
163. Sun, H.; Zhang, P. The Role of Moderating Factors in User Technology Acceptance. *Int. J. Hum.-Comput. Stud.* **2006**, *64*, 53–78. [CrossRef]
164. FakhrHosseini, S.; Chan, K.; Lee, C.; Jeon, M.; Son, H.; Rudnik, J.; Coughlin, J. User Adoption of Intelligent Environments: A Review of Technology Adoption Models, Challenges, and Prospects. *Int. J. Hum.–Comput. Interact.* **2024**, *40*, 986–998. [CrossRef]

165. Ilieva, G.; Yankova, T.; Ruseva, M.; Dzhabarova, Y.; Zhekova, V.; Klisarova-Belcheva, S.; Mollova, T.; Dimitrov, A. Factors Influencing User Perception and Adoption of E-Government Services. *Adm. Sci.* **2024**, *14*, 54. [CrossRef]
166. Leesakul, N.; Oostveen, A.M.; Eimontaite, I.; Wilson, M.L.; Hyde, R. Workplace 4.0: Exploring the Implications of Technology Adoption in Digital Manufacturing on a Sustainable Workforce. *Sustainability* **2022**, *14*, 3311. [CrossRef]
167. Davis, F.D. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.* **1989**, *13*, 319–340. [CrossRef]
168. Shin, D. Understanding User Acceptance of DMB in South Korea Using the Modified Technology Acceptance Model. *Int. J. Hum. Comput. Interact.* **2009**, *25*, 173–198. [CrossRef]
169. Bagozzi, R. The Legacy of the Technology Acceptance Model and a Proposal for a Paradigm Shift. *J. AIS* **2007**, *8*, 243–254. [CrossRef]
170. Tejero Gimeno, P.; Pastor, G.; Choliz, M. On the concept and measurement of driver drowsiness, fatigue and inattention: Implications for countermeasures. *Int. J. Veh. Des.* **2006**, *42*, 67–86. [CrossRef]
171. Breazeal, C. Toward social robots. *Robot. Auton. Syst.* **2003**, *42*, 167–175. [CrossRef]
172. Selwyn, N. The use of computer technology in university teaching and learning: A critical perspective. *J. Comput. Assist. Learn.* **2007**, *23*, 83–94. [CrossRef]
173. Nguyen, A.; Kremantzis, M.D.; Essien, A.; Petrounias, I.; Hosseini, S. Enhancing Student Engagement Through Artificial Intelligence (AI): Understanding the Basics, Opportunities, and Challenges. *J. Univ. Teach. Learn. Pract.* **2024**, *21*. [CrossRef]
174. Kamel, S.; Dahawy, K. Perception and/or Individual Difference: What Affects the Acceptance of New Technology? In Proceedings of the International Business Information Management Association Conference (IBIMA) on the Internet and Information Technology in Modern Organizations, Cairo, Egypt, 13–15 December 2005. [CrossRef]
175. Xu, H.; Teo, H.; Tan, B.; Agarwal, R. Research Note—Effects of Individual Self-Protection, Industry Self-Regulation, and Government Regulation on Privacy Concerns: A Study of Location-Based Services. *Inf. Syst. Res.* **2012**, *23*, 1342–1363. [CrossRef]
176. Lee, C.; Coughlin, J.F. PERSPECTIVE: Older adults' adoption of technology: An integrated approach to identifying determinants and barriers. *J. Prod. Innov. Manag.* **2015**, *32*, 747–759. [CrossRef]
177. Culnan, M.; Armstrong, P. Information Privacy Concerns, Procedural Fairness and Impersonal Trust: An Empirical Investigation. *Organ. Sci.* **1998**, *10*, 104–115. [CrossRef]
178. Friedman, B.; Nissenbaum, H. Bias in computer systems. *ACM Trans. Inf. Syst.* **1996**, *14*, 330–347. [CrossRef]
179. Voigt, P.; Bussche, A. The EU General Data Protection Regulation (GDPR). In *A Practical Guide*; Springer: Cham, Switzerland, 2017. [CrossRef]
180. Chau, P.Y.K. An Empirical Assessment of a Modified Technology Acceptance Model. *J. Manag. Inf. Syst.* **1996**, *13*, 185–204. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.