Towards Human-Compatible AI for Well-being by Integrating Physiological Viewpoint With Machine Learning Viewpoint

Keiki Takadama[1], Daiki Shintani[2]

[1]The University of Tokyo, Japan
[2]The University of Electro-Communications, Japan
takadama@g.ecc.u-tokyo.ac.jp, shindaiki@cas.lab.uec.ac.jp

Abstract

This paper focuses on "human-compatible AI" which aligns with human values and remains under human control to prevent unintended and harmful consequences, and discusses it to develop human-compatible AI for well-being. For this issue, this paper proposes the human-compatible AI for a sleep as one of the human-compatible AI for well-being, which is designed to have the functions of (1) checking how the estimated sleep stage (corresponding to suggestions to users) follows the biological rhythms which determine their health conditions (corresponding to human values) and (2) modifying the estimated sleep stage according to their biological rhythms. To investigate an importance of the proposed approach, this paper applies it into the sleep stage estimation based on the acceleration sensor data. Through the human subject experiment, the following implications have been revealed: (1) it is dangerous to simply employ machine learning (*i.e.*, Random Forest in this research) for the sleep stage estimation because the sleep stage is artificially estimated without following the ultradian rhythm which are generally found in humans; and (2) it is important to integrate the physiological characteristic (*i.e.*, the ultradian rhythm) with machine learning for the sleep stage estimation because such an integration can estimate the sleep stage that follow the ultradian rhythm.

Introduction

On April 2nd, 2024, WHO (Would Health Organization) announced to release S.A.R.A.H. (a Smart AI Resource Assistant for Health), which is the digital health promoter prototype with enhanced empathetic response powered by generative artificial intelligence (AI) (WHO 2020a). S.A.R.A.H. is developed as the AI health information avatar which aims at providing information across major health topics, including healthy habits and mental health, through a conversation between the S.A.R.A.H. avatar and a user (WHO 2020b). In particular, S.A.R.A.H. can provide tips to destress, eat right, quit tobacco and e-cigarettes, as well as give information on several other health topics, in order to help users to manage and optimize his/her health.

However, S.A.R.A.H. cannot guarantee to always provide the correct or appropriate information/suggestions for the users who ask for advices of their good health, *i.e.*, it may provide the wrong or inappropriate answers to users. This is a serious problem because such information/suggestions affect their health even though humans easily believe the outputs generated by AI that contains false or misleading information presented as fact (known as "hallucination") or humans easily rely only on the first AI recommendation and not to explore alternatives (known as "anchoring bias"). More importantly, this problem is very hard to be avoided because the definition of the "good health" is ambiguous, which makes it difficult for users to judge whether the outputs generated by AI is correct/appropriate or wrong/inappropriate. Furthermore, even if such information/suggestions are useful, they cannot guarantee to always derive good health "every day," *e.g.*, they may be useful most day but not in a certain day, or they may be useful in a certain day but not in other days. Since they are partially correct or appropriate, such answers also resulting in unconscious believe or reliance of them. It goes without saying that this situation gets worse when a quality of the answers increases by becoming the current AI to superintelligent through a self-learning of data in the world.

For this issue, Russell proposed "human-compatible AI" (Russell 2019) which aligns with human values and remains under human control to prevent unintended and harmful consequences. As the main point of human-compatible AI, an objective of AI should remain *uncertain* unlike an objective of an agent in the conventional AI is rigid and certain. This is because such uncertainty makes it hard for AI to execute concrete actions according to the uncertain objective, which gives a chance for AI to learn human values by ob-

serving human behaviors. This process contributes to preventing misunderstandings of human values, which should be maximized as the true objective of AI.

From this viewpoint, the objective of deriving good health is *uncertain* because we do not clearly know how to achieve it. This uncertainty is a good characteristic to develop human-compatible AI for well-being. However, both S.A.R.A.H. and the current generative AI such as Large Language Models (LLMs) (Wei et al. 2022) do not understand human values but simply predicts the next tokens (which is a minimum unit of sentences) estimated by the transformer technology (Vaswani 2017). This means that S.A.R.A.H. and the generative AI always provides some suggestions when asking for advices of their good health without checking whether the given suggestions satisfy human values. This is the significant problem because appropriate suggestions generally depend on human values on good health (*e.g.*, a sleep is a concern of some persons, while nutrition of meals is a concern of other persons). For this issue, human-compatible AI for well-being should design S.A.R.A.H. or the generative AI to have the functions of (1) checking how the suggestions satisfy human values of users and (2) modifying suggestions according to their values.

Towards such human-compatible AI, this paper starts to address a *sleep stage estimation* from biological data such as heartrate, respiration, and body movement by using machine learning technologies, while checking how the estimated sleep stage follows the "biological rhythms". What should be noted, here, is that the *biological rhythms* are not *human values* but the output of machine learning should follow the biological rhythms when providing it as the suggestion for human health because the biological rhythms affect our health conditions. For example, health condition is generally good/bad when an "ultradian rhythm" (*i.e.*, approximately 90 minutes rhythm) as one of biological rhythms of humans is stable/unstable. Since good health (caused by the stable ultradian rhythm) increases human values, this paper roughly regards "a follow of a biological rhythm" as "a satisfaction of human values." From this viewpoint, the sleep stage estimated by machine learning without checking how the estimated sleep stage follows a biological rhythm is dangerous because it may provide wrong/inappropriate message. To overcome this issue, this paper proposes the human-compatible AI for a sleep as one of the human-compatible AI for well-being, which is designed to have the functions of (1) checking how the estimated sleep stage (corresponding to suggestions to users) follows the biological rhythm (corresponding to human values) and (2) modifying the estimated sleep stage according to their biological rhythm.

This paper is organized as follows. The next section introduces human-compatible AI. After Section 3 explains the biological rhythm and sleep stage, Section 4 conducts the human subject experiments and discusses the results in Section 5. Finally, our conclusion is given in Section 6.

Human-Compatible AI

Conventional AI

The conventional AI is generally executed according to given human-specified goals, but such goals may not reflect an intention of humans. For this issue, Russell in his TED talk introduced "the off-switch problem", which argues whether a machine (AI) let us switch it off, using the example of the robot which is asked to fetch a coffee as the given objective (Russell 2017). The standard robot may think as follows.

Robot: "I must fetch a coffee."
Robot: "I can't fetch a coffee when I'm dead."
Robot: "Therefore, I must disable my off-switch."

Such an unexpected outcome comes from the certain objective, meaning that any human values are not included in the objectives.

Human-Compatible AI and Three Principles

To overcome the problem of the conventional AI which is based on certain objective, Russell proposed "human-compatible AI" (Russell 2019) which is based on the uncertain objective. Concretely, human-compatible AI would have the true objective remain uncertain, which encourages AI for cooperation and communication with humans to increase certainty about it by gaining more information about human values. Such human-compatible AI provides us the provably beneficial machines that focus on deference to humans. For this issue, the following three principles are important (note that the term "preference" is employed in the original definition instead of "value" which is employed in the TED talk).

1. The machine's only objective is to maximize the realization of human values.
2. The machine is initially uncertain about what those values are.
3. The ultimate source of information about human values is human behavior.

Ultradian Rhythm and Sleep Stage

Ultradian Rhythm

The biological rhythm in human body is composed of a lot of rhythms, e.g., a month rhythm as a circalunar rhythm, a weak rhythm as a circaseptan rhythm, 24 hours rhythm as a circadian rhythm, and 90 minutes rhythm as an ultradian rhythm). Although all of rhythms affect health conditions, an ultradian rhythm is focused on because of the following reasons: (1) the health condition of one day is generally useful than that of week and month, which means that the circadian or ultradian rhythm become the candidates; (2) the

ultradian rhythm can be found during sleep (*i.e.*, the deep and light sleep are continuously repeated in a cycle from 60 to 120 minutes) while the circadian rhythm cannot be found during sleep, and the nighttime biological data is more stable due to less activity than the daytime data, which contributes to decreasing the difficulty of estimating an ultradian rhythm from the sleep stage.

Sleep Stage

The sleep stage is an indicator of the depth of sleep, which is defined by the R&K method (Rechtschaffen and Kales 1968) based on electrooculogram (EEG), electromyogram (EMG), and electrooculogram (EOG) acquired during sleep. Concretely, the sleep stages are divided into five stages, *i.e.*, the wake stage, the REM sleep stage, and the Non-REM sleep stages 1, 2, and 3, represented by WAKE, REM, NREM1, NREM2, and NREM3, respectively. As shown in Figure 1 where the horizontal axis indicates the sleep time in a bed while the vertical axis indicates the sleep stage, WAKE is the lightest sleep while NREM3 is the deepest sleep. Note that NREM3 and NREM4 are merged into one stage (as NREM3) according to the American Academy of Sleep Medicine (AASM) scoring manual (Berry et al. 2012). The five sleep stages are determined with an interval of the 1 epoch (*i.e.*, the 30 seconds) by the human experts.

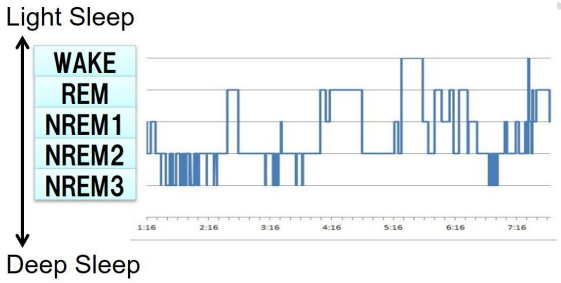

Figure 1: Sleep stage

Sleep Stage Estimation Based on Ultradian Rhythm

Our previous research (Shintani et al. 2024a; 2024b) developed the sleep stage estimation for the simple acceleration sensor. Unlike many conventional methods of the sleep stage estimations based on an acceleration sensor (Boe et al. 2019; Sundararajan et al. 2021; Gu et al. 2014), our proposed method does not simply employ machine learning or improves it, but indirectly taking account of the ultradian rhythm when estimating the sleep stage by machine learning. For this issue, our proposed method starts to (i) determines the epochs of WAKE, followed by determining the epochs of NREM3 (Shintani et al. 2024a) and REM (Shintani et al. 2024b) in turn, which means that the sleep stage is estimated under the priority of WAKE > NREM3 > REM; and then (2) the sleep stage of the remaining epochs is determined as NREM1 and 2. Furthermore, REM and NREM3 are estimated by checking both of them follows the ultradian rhythm. Concretely, the estimated REM and NREM3 which follow the ultradian rhythm are remained while those which do not follow it are changed to other sleep stages, *i.e.*, NREM3 may be changed to REM or NREM1 and 2 while REM may be changed to NREM 1 and 2 under the priority of WAKE > NREM3 > REM > NREM 1 and 2. What should be noted here is that (i) the approach of "firstly" determining WAKE and REM as the light sleep and NREM3 as the deep sleep contributes to emphasizing the light and deep sleep, *i.e.*, the ultradian rhythm; and (ii) the approach of estimating only REM and NREM3 which follow the ultradian rhythm contributes to taking account of the ultradian rhythm.

Human Subject Experiment

Experimental Setup

To investigate an influence of taking account of the ultradian rhythm when estimating the sleep stage, the human subject experiment was conducted with the approval of the ethics community of Ota General Hospital for this study in agreement with Helsinki's declaration. Through this experiment, the 35 whole nights sensor data of the healthy human subjects were obtained, where their ages are ranged from 20's to 60's including both genders.

As the acceleration sensor, this paper employs the coin-shaped sensor (BRAIN SLEEP COIN, Brain Sleep Co. Ltd.) attached to a nightwear around waist, meaning that the sleep stage can be estimated without connecting any devices to human's body. The sleep stage is estimated from the biological vibration data acquired from the coin-shaped sensor, which measures the following data: (i) the norm data of acceleration with a sampling rate of 10 Hz, (ii) the z-axis data of acceleration with a sampling rate of 1 Hz (1: face up, 0: sideways, -1: face down), and (iii) the temperature with a sampling rate of 1 Hz. Among these data, the norm and z-axis of acceleration are employed in the experiment. In addition to the "estimated" sleep stage based on the biological vibration data of the coin-shaped sensor, the "correct" sleep stage was obtained through the PSG (polysomnography) test based on the EEG, EOG, and EMG data.

Evaluation Criterion and Parameter Setting

The experiment employs the leave-one-out cross-validation, where the model is trained on the 34 nights data and evaluated on the other night. As the evaluation criterion, an accuracy between the estimated sleep stage and the correct sleep stage of the PSG test is employed.

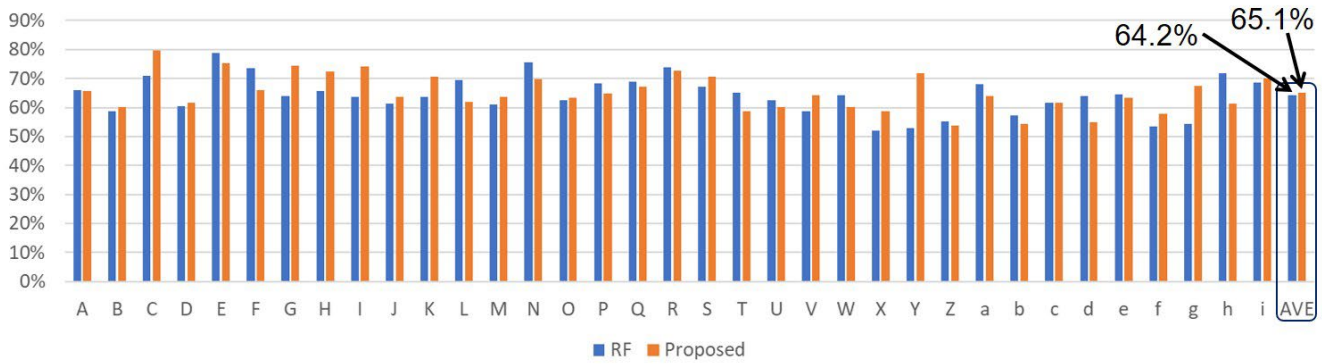

Figure 2: Accuracy of the sleep stages estimated by RF and the proposed method

To clarify an influence of the ultradian rhythm, the sleep stage estimated by our previous method is compared with that by Random Forest (RF) (Breiman, 2001) as one of the major machine learning methods. RF is an ensemble learning method composed of the multiple decision trees as a weak classifier and determines the output (*i.e.*, the classification result) by the majority vote of the classification results of the decision trees. The parameters of RF were set as follows: (i) the maximum depth of the decision tree is 10; and (ii) the number of decision trees is 100. Note that the four stages (WAKE, REM, NREM1 and 2, NREM3) where NREM1 is merged with NREM2 are employed instead of the five stages in this experiment because of the following reasons: (1) the ratio of NREM1 is smaller than that of NREM2; and (2) its merger contributes to increasing the accuracy of the WAKE, REM and NREM3 in the four stages (i.e., the three stages estimation out of the four stages) in comparison with that in the five stages (i.e., the three stages estimation out of the five stages). This means that the merging NREM1 with NREM2 gives an advantage to RF in terms of estimating the sleep stage which shows the ultradian rhythm. For the training of RF, all data of WAKE, REM and NREM3 are employed as the training data while the 50% data of NREM 1 and 2 which are randomly selected from all data of NREM 1 and 2 are employed as the training data. This is because an amount of the data of NREM 1 and 2 is quite larger than that of WAKE, REM and NREM3 in the original data, *i.e.*, the ratio of the data of NREM 1 and 2 is more than 50 %.

Experimental Result

Figure 2 shows the accuracy of the sleep stages estimated by RF and the proposed method, where the horizontal axis indicates the 35 subjects with their average while the vertical axis indicates the accuracy of the sleep stage. In this figure, the blue and orange bars indicate the accuracy of the sleep stages estimated by RF and that by the proposed method, respectively. When focusing on the averaged accuracy, both accuracies of the sleep stages estimated by RF (64.2%) and

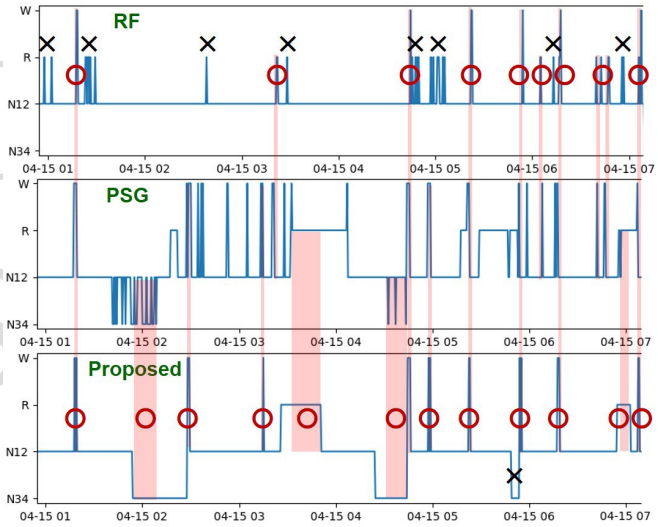

Figure 3: Sleep stages of Subject A by RF, PSG, and the proposed method)

the proposed method (65.1%) are very similar. This tendency can be found in the accuracy in each subject, even though the accuracy of RF is higher/lower than that of the proposed method in some subjects. From the viewpoint of statistical analysis, Wilcoxon signed-rank test shows that the p values of the averaged accuracies of the sleep stages estimated by RF and the proposed method is 0.69, meaning that no significant difference is found.

To investigate an influence of taking account of the ultradian rhythm, Figure 3 shows the sleep stages of the subject A, where the horizontal axis indicates the sleep time in a bed while the vertical axis indicates the sleep stage. In detail, the upper, middle, and lower graphs indicate the sleep stages of RF, PSG, and the proposed method, respectively. The red circle and the black cross mark indicate the correct and wrong estimation of WAKE, REM, and NREM3, respectively. The light red bars are the period of the correct estimation of WAKE, REM, and NREM3. Here, the sleep stage of the subject A is employed because the accuracies of RF

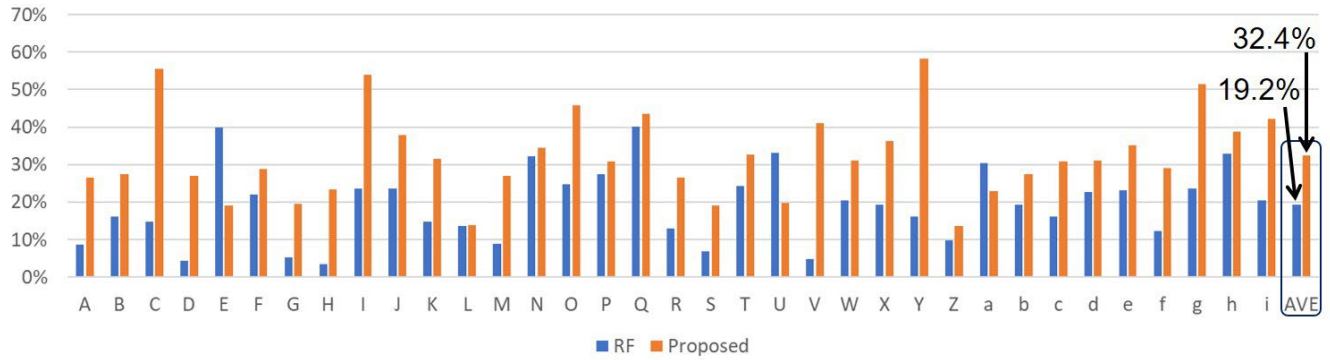

Figure 4: Weighted kappa of the sleep stages estimated by RF and the proposed method

(65.93%) and the proposed (65.69%) are mostly the same from Figure 2. However, Figure 3 clearly shows the different sleep stages of RF and the proposed method. Concretely, NREM3 are not found in the sleep stage of RF but found in that of the proposed method, and many WAKEs and REMs are wrongly estimated by RF while all WAKEs and some REMs are correctly estimated by the proposed method. These results suggest that the sleep stage of RF does not follow the ultradian rhythm due to no cycle of the light (*i.e.*, (the correct) WAKE and REM) and deep (*i.e.*, NREM3) sleep while that of the proposed method roughly follow the ultradian rhythm due to a cycle of the light and deep sleep even though the periods of the light and deep sleep are not perfectly the same.

Discussions

Why Is Ultradian Rhythm Needed To Estimate Sleep Stage?

Even though the sleep stage of RF is different that the proposed method as shown in Figure 3, both accuracies of the sleep stage are mostly the same, because of the following reasons: (1) RF aims to minimize a difference of the sleep stage of RF and PSG, which promotes RF to learn NREM1 and 2 preferentially because the ratio of NREM 1 and 2 is larger than that of others (*i.e.*, its ratio is more than 50% of a sleep time). From this reason, the sleep stage estimated by RF is mostly NREM1 and 2 without estimating REM and NREM3; (2) The proposed method takes account of the ultradian rhythm when estimating NREM3 and REM, and it estimates the sleep stage under the priority of WAKE > NREM3 > REM > NREM1 and 2, which contributes to emphasizing the light sleep (*i.e.*, WAKE and REM) and deep (*i.e.*, NREM3) sleep. From this reason, the sleep stage estimated by the proposed method roughly follows the ultradian rhythm.

Evaluating Sleep Stage From Weighted Kappa

Since it is difficult to show the difference of RF and the proposed method (*i.e.*, RF mainly estimates NREM1 and 2, while the proposed method estimates the four stages) from the viewpoint of the accuracy of the sleep stage, the weighted kappa is focused on because its criterion can evaluate a degree of the difference between the sleep stages (*e.g.*, the difference between WAKE and NREM3 is larger than that between REM and NREM1 and 2). Employing this criterion, Figure 4 shows the weighted kappa of the sleep stages estimated by RF and the proposed method, where the horizontal axis indicates the 35 subjects with their average while the vertical axis indicates the accuracy of the sleep stage. In this figure, the blue and orange bars indicate the weighted kappa of the sleep stages estimated by RF and that by the proposed method, respectively. When focusing on the averaged weighted kappa, the value of the proposed method (32.4%) is larger than that of RF (19.2%). This tendency can be found in the weighted kappa in each subject except for the subjects E, U, and a. From the viewpoint of statistical analysis, Wilcoxon signed-rank test shows that the p values of the averaged weighted kappa of the sleep stages estimated by RF and the proposed method is 2.04×10^{-6}, meaning that the significant difference is found.

This analysis has revealed the following implications: (1) it is dangerous to simply employ machine learning (*i.e.*, RF in this research) for the sleep stage estimation because the sleep stage is artificially estimated without following the ultradian rhythm which are generally found in humans; and (2) it is important to integrate the physiological characteristic (*i.e.*, the ultradian rhythm) with machine learning for the sleep stage estimation because such an integration can estimate the sleep stage that follow the ultradian rhythm.

Conclusion

This paper focused on human-compatible AI and discussed it to develop human-compatible AI for well-being. For this

issue, this paper focused on an ultradian rhythm and proposed the human-compatible AI for a sleep as one of the human-compatible AI for well-being, which is designed to have the functions of (1) checking how the estimated sleep stage follows the biological rhythms which determine their health conditions and (2) modifying the estimated sleep stage according to their biological rhythms. To investigate an importance of the proposed approach, this paper applies it into the sleep stage estimation based on the acceleration sensor data. The human subject experiment has revealed the following implications: (1) it is dangerous to simply employ machine learning (*i.e.*, RF in this research) because machine learning artificially provides the output which does not follow the physiological characteristics which are generally found in humans; and (2) it is important to integrate the physiological characteristic (*i.e.*, the ultradian rhythm) with machine learning because such an integration can provides the output which follows the physiological characteristics. rhythm.

What should be noticed here is that this paper takes just one step for developing human-compatible AI for well-being, therefore the next step must be pursued in the near future in addition to (1) applying the proposed approach to S.A.R.A.H. and the generative AI; and (2) exploring the criterion of measuring the uncertainty from viewpoint of the ultradian rhythm.

Acknowledgments

The research reported here was supported in part by a Grant-in-Aid for Scientific Research (A) (22H00547) and Grant-in-Aid for Challenging Exploratory Research (24K22334) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan.

References

Berry, R. B., Brooks, R., Gamaldo, C. E., Harding, S. M., Marcus, C., and Vaughn, B. V. 2012. "The AASM manual for the scoring of sleep and associated events Rules, Terminology and Technical Specifications," *American Academy of Sleep Medic*ine, Vol. 176.

Boe, A. J., McGee Koch, L. L., O'Brien, M. K., Shawen, N., Rogers, J. A., Lieber, R. L., Reid, K. J., and Zee, P. C.; and Jayaraman, A. 2019. "Automating sleep stage classification using wireless, wearable sensors," *NPJ digital medicine*, Vol 2. No. 131.

Breiman, L. 2001. "Random forests", *Machine learning*, Vol. 45, No. 1, pp. 5-32.

Gu, W., Yang, Z., Shangguan, L., Sun, W., Jin, K., and Liu, Y. 2014. "Intelligent sleep stage mining service with smartphones," *The 2014 ACM international Joint Conference on pervasive and ubiquitous Computing*, pp. 649-660.

Rechtschaffen, A.; and Kales, A. 1968. *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects*, Public Health Service, US Government Printing Office.

Russell, S. J. 2017. "3 principles for creating safer AI," TED talk, https://www.ted.com/talks/stuart_russell_3_principles_for_creating_safer_ai.

Russell, S. J. 2019. "Human Compatible: Artificial Intelligence and the Problem of Control," Viking publisher.

Shintani, D., Nakari, I., Washizaki, S., and Takadama, K. 2024a, "NREM3 Sleep Stage Estimation Based on Accelerometer by Body Movement Count and Biological Rhythms," *The AAAI 2024 Spring Symposia, Impact of GenAI on Social and Individual Well-being*, Vol. 3, No. 1, pp. 405-411.

Shintani, D., Nakari, I., Washizaki, S., and Takadama, K. 2024b, "REM Estimation Based on Accelerometer by Excluding Other Stages and Two-Scale Smoothing," The 46th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC2024).

Sundararajan, K., Georgievska, S., Te Lindert, B. H. W., Gehrman, P. R., Ramautar, J., Mazzotti, D. R., Sabia, S., Weedon, M. N., van Someren, E. J. W., Ridder, L., Wang, J., van Hees, V. T. 2021. "Sleep classification from wrist-worn accelerometer data using random forests," *Scientific reports*, Vol 11, No. 24.

Vaswani, A. Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. 2017 "Attention Is All You Need", Neural Information Processing Systems, pp. 6000-6010.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, Donald., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean J., and Fedus, W., 2022, "Emergent Abilities of Large Language Models", Transactions on Machine Learning Research (TMLR).

Would Health Organization, 2024a. "S.A.R.A.H, Smart AI Resource Assistant for Health" https://www.who.int/campaigns/s-a-r-a-h.

Would Health Organization), 2024b. "WHO unveils a digital health promoter harnessing generative AI for public health," https://www.who.int/news/item/02-04-2024-who-unveils-a-digital-health-promoter-harnessing-generative-ai-for-public-health.