Yonatan Belinkov is an Associate Professor at the Faculty of Computer Science at the Technion. He received his PhD from MIT in 2018, followed by post-doctoral research at Harvard, where he was a Mind Brain Behavior Post-doctoral Fellow. His research interests are in Artificial Intelligence and Machine Learning, especially interpreting the internal mechanisms of AI models, assessing their robustness and safety, and improving their controllability. He is also broadly interested in multi-agent communication and in AI for the Natural Sciences. Yonatan is a former Azrieli Faculty Fellow and has been awarded the Krill Prize from the Wolf Foundation .