# Human-Instrument Co-adaptation

Let me describe to you a long-lasting frustration and a new hope.

The current channels of musical expression are like thin tubes that musical ideas have to squeeze through. A Digital Audio Workstation (DAW) is expressive but not real-time. A trumpet is real-time but not nearly as expressive. How can a machine extract humans' internal music imagination, so that a genius like Bach could compose the four parts of a four-part fugue simultaneously; so that a non-expert could improvise music without mastering any instrument or DAW?

Allow me to show you my answer: a dynamically scaffolded multi-modal *co-adaptation* between a human and her bespoke instrument through interactive machine learning. Picture an instrument that translates the human’s hand gesture, body motions, breathing, micro facial expressions, tongue movement details, muscle activations, and EEG signal into the music she is imagining. The instrument uses an ensemble of variational neural networks supervised by parallel data generated when she listens to or improvises music. Two training techniques foster better generalization from limited training data: 1) KL divergence loss ensures *low-frequency* *continuity* of the mapping; 2) Cycle consistency makes the entire output space *reachable*. Additionally, the human plays the instrument, marks undesirable behaviors, and scores the various ensemble learners in offline review sessions.

Furthermore, the human learns the instrument! Contrary to a static decoding task where the ground truth is passive, in co-adaptation the ground truth also moves towards the decoder. Concretely, when the human sees something the instrument does, she latches onto that and has some low-dimensional control over the instrument. She tries exposing different features for the instrument to learn. The human-instrument bi-directional learning never ends and the dimensionality of control increases. Moreover, the instrument applies haptic guidance to train the human. Specifically, the human selects a piece she wants to play, and the haptic ground truth is computed by the opposite encoder in charge of the cycle consistency. Using haptic guidance, the human can also intuitively playback old training data and tell the instrument to “forget” outdated ones.

Here is the big picture in my eyes. Mind-reading the human would offer perfect expressivity, but unintrusive readings like EEG are unassailably noisy – an unreliable mapping. A piano translates finger motions to music, but is not nearly as expressive – a collapsed mapping. Co-adaptation gradually finds a mapping where the human can interface with the instrument at maximal information throughput. With many users, we can summarize several “pruned / principal instruments” for beginners to fork and jump start the training.

We will finally be free! Non-experts will be able to jam together in symphony. Musicians will be able to improvise multi-part novel-timbre music while hearing it in real time!

## More Points (that I can think of so far)

* The initialization of the mapping can be copied from an existing instrument.
* We can go from a *generator* that decodes random noise to music, to a *controlled generator* that decodes some noise in junction with human input, and finally to an *instrument* that decodes the human.
* The [above-mentioned techniques](#techniques) (that ensure mapping continuity and output space reachability) not only regularize the instrument via training, but also expose problems with the human’s envisioned mapping via the instrument’s learning behaviors, resulting in a human-machine collaborated design process.
* Online learning enables the human to supply training data in response to the instrument’s mistakes [].
* In turn, the machine actively asks for training data. “How would you play this music?” “What would this sound like? [applies haptic guidance]” Asking good questions requires good unsupervised learning.
* Another training method is to let the human sight-play a musical score on the instrument. Here both the human and the instrument learn at the same time. For the human, the (musical score, instrument) pair is the ground truth that supervises her playing. For the instrument, the (musical score, human) pair is the ground truth that supervises its parsing. It is analogous to cue-based BCI learning defined in [].
  + Also, a muted variation of sight-play training may better approximate the improvisation environment, fighting the [domain danger](#danger). The instrument use eye tracking to align human input with the score. Afterwards, the human marks her own mistakes so they don’t become training data, which will be easy with haptic playback.
* Automatic discovery of redundancy and query the user to cut complexity.
  + Redundancy can be defined as either (A) how well can a smaller network learn to replicate the current network, or (B) how model accuracy responds to regularization strength.
* Automatic discovery of problematic / conflicting training data.
  + For example, (A) as we strengthen regularization, which training datapoints incur high marginal loss? (B) masking which training datapoints minimizes the regularization loss of a model trained to equilibrium?
  + Even better, we want to show the user a pair of contradictory datapoints. Cluster training data into a tree where each node has a contradiction index. Find the maximum contradiction node. Represent the a subnode with generated summary, if it is not a leave yet.
* Dampen the instrument’s learning, so that the human is not dealing with a stranger every batch [].
* The machine learning (ML) model.
  + It is *not* human-in-the-loop reinforcement learning (RL) since rewards are not sparse and feedbacks are directional.
  + It is more like a combination of VAE and CycleGAN.
  + Compared to the logic of the Wekinator, it involves a two-way mapping, allowing the machine to figure things out on its own, query the human efficiently, and even guide the human with its inductive biases.
* Are humans capable of such rapid information output?
  + Novel human modality discovery: see [my blog](/#/blog/new_modalities). Also,
  + I invite you to imagine a world where the piano did not exist. You asked people in that world to imagine what it would be like to play the piano. They would tell you it was impossible! Such unintuitive and precise skills must be too difficult for humans to acquire. From an objective lens, playing the piano is a truly novel way for the human to output information. Natural selection did not make this happen. It is the plasticity of human capability that makes it possible. We underestimate our potential to interact. Our output capabilities are vastly unused.
* Domain danger.
  + When the human trains the instrument with different methods and in different environments, the data will be in different domains, making few-shot learning more difficult.
  + EEG has non-stationarities [].
  + Particularly, “users exhibit a different mental state during offline calibration and online feedback” []. In my case the online feedback is not a problem but the entire goal. Hence the question is how to make the offline calibration more useful for the online feedback. [] applies online unsupervised learning to normalize feedback-time input. In my case, since music control should be precise, domain adaptation needs to happen fast.
* Three levels of HCI + ML:
  + Interactive machine learning (**IML**) originally proposed by []: Online learning allows the human to provide more effective training data. The human’s goal does not change.
  + **Co-adaptation**. The human changes her desired mapping during the process [fiebr, fails]. On the bright side, the human avoids what the machine is bad at and focuses on what the machine is good at, increasing model accuracy. More importantly, interesting responses from the machine give the human new ideas, making the machine no longer a passive tool but a counterparty in the creative duo. But of course, the moving target poses new challenges to the training of the machine [mv tgt].
  + **Bi-directional learning**. While training, the machine teaches a skill to the human. The bi-directional learning drastically redefines the learning problem and further promotes the role of the machine in the interaction. In the context of my proposal, recent advancement in haptic guidance [] is a core prerequisite.
* Correct machine mistake (provide ground truth) at low cost with the machine’s help.
  + The main reason I want to use ensemble learning is to provide alternative parses of the human’s playing, so that the human can easily select a nice ground truth among candidates.
  + In CueTIP [], when the machine makes a mistake, the human provides partial constraints and let the machine re-classify until correct results are achieved. It will be nice if I can think of some ways the human can communicate musical constraints with the machine to implement a similar tool.

## Related Works