FaceShift

Galip Ümit Yolcu

Politics of Machines - Studio Class - WS 21/22

Initial topic: Politics of Technology and Patterns
Current topic: Politics of Information Technologies and Meaning

informationprocessing
meaningmaking
noise

8 March 2022

Contents

Contents ii

Chapter 1 Context: Historical and Philosophical Background of Information Processing Technologies 1
1.1 Introduction . . . 1
1.2 Information Theory . . . 2
1.3 Cybernetics . . . 4
1.3.1 Signals and Systems . . . 4
1.3.2 Feedback Systems . . . 4
1.3.3 Noise . . . 5
1.3.4 Communication and Control . . . 6
1.4 Learning Machines . . . 6
1.4.1 Just another box, but black . . . 7
1.4.2 The paradox of designing an intelligent system . . . 8

Chapter 2 Urgency: Political Implications of Intelligent Systems 11
2.1 Meaning Making . . . 11
2.2 Dirty Data . . . 12
2.3 Automated Decision Making . . . 12
2.4 Gamification - Machinization . . . 13

Chapter 3 Intervention: Cancelling Noise Cancelling 14

Chapter 4 Design Project: FaceShift 15
4.1 Motivation . . . 15
4.1.1 Exploring clashing semantic spaces . . . 15
4.1.2 Purposefulness/Intentionality Recognition . . . 15
4.2 Technical Background . . . 16
4.2.1 Dimensionality Reduction . . . 16
4.2.2 Data Spaces . . . 16
4.3 Design . . . 17
4.3.1 Autoencoders . . . 17
4.3.2 FaceShift . . . 19
4.4 Resources and References . . . 20

Chapter 5 Reflection and Conclusion 22

Speculative Concepts 25

Course Evaluation 26

References **27**

CHAPTER 1

Context: Historical and Philosophical Background of Information Processing Technologies

---

1.1 Introduction

Information processing is a loaded term. While it is generally understood as

> "the acquisition, recording, organization, retrieval, display, and dissemination of information." [**Slamecka, 2018**]

it is also associated almost always with computers or things that are thought to be computing machines. The Encyclopedia of Computer Science summarizes the concept as

> "Information processing might, not inaccurately, be defined as "what computers do." In fact, the broadest professional organization concerned with computer science is named the International Federation for Information Processing."[**Simon, 2003**]

In fact, this concept has attracted many sciences as a general framework, gave rise to new sciences, and to branches under existing ones. Some examples are Computational Psychology, Computational Linguistics, Computational Biology, Machine Learning, Cognitive Science, Data Science.

Moreover, these approaches seem to all have been primarily influenced by three frameworks of modeling coming from a mathematical background: computing machinery, information theory and cybernetics.[**Heims, 1989**]

Claude E. Shannon, the leading founder of information theory has admitted one of the reasons for the theory's appeal is its compatibility with and connections to the other two approaches, and actively tried to warn scientists of the danger of being convinced by this theory, despite its narrow range of applicability:

> "Although this wave of popularity is certainly pleasant and exciting for those of us working in the field, it carries at the same time an element of danger. While we feel that information theory is indeed a valuable tool in providing fundamental insights into the nature of communication problems and will continue to grow in importance, it is certainly no panacea for the communication engineer or, a fortiori, for anyone else. Seldom do more than a few of nature's secrets give way at one time. It will be all too easy for our somewhat artificial prosperity to collapse overnight when it is realized

> that the use of a few exciting words like *information*, *entropy*, *redundancy*, do not solve all our problems."[**Shannon, 1956**]

Cyberneticians, on the other hand, were much more confident on the wide range of applicability of their model. In fact, cyberneticians seem to think of their method as a new science that tries to encapsulate all others, or as a new way of thinking about physics, and by consequence, the world. [**Dupuy, 2009**][**Wiener, 1989a**]

Meanwhile, there has also been another narrative around Cybernetics that is too fun to not mention:

> "**Cybernetics:** a reactionary pseudoscience that appeared in the U.S.A. after World War II and also spread through other capitalist countries. Cybernetics clearly reflects one of the basic features of the bourgeois worldview—its inhumanity, striving to transform workers into an extension of the machine, into a tool of production, and an instrument of war. At the same time, for cybernetics an imperialistic utopia is characteristic—replacing living, thinking man, fighting for his interests, by a machine, both in industry and in war. The instigators of a new world war use cybernetics in their dirty, practical affairs."[**Peters, 2012**]

The philosophical implications of cybernetics taken literally, mostly concern philosophy of mind, language and science. However, we think some of their basic definitions and assumptions give a way of interpreting immediate political problems raised by the use of machines, automated systems, and especially artificial intelligence today, as will be presented in chapter 2. Therefore, we do not mention connections to old philosophical problems and modern physics[**Wiener, 1989a**] and focus on providing an intuitive explanation of the general framework and the fundamental conceptions that are relevant.

1.2 Information Theory

Information theory was put forth by Claude Shannon and Warren Weaver in their seminal 1948 paper: A Mathematical Theory of Communication.

They developed the theory during the Second World War. The paper neatly describes the engineering problem in question:

> "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design."[**Shannon and Weaver, 1948**]

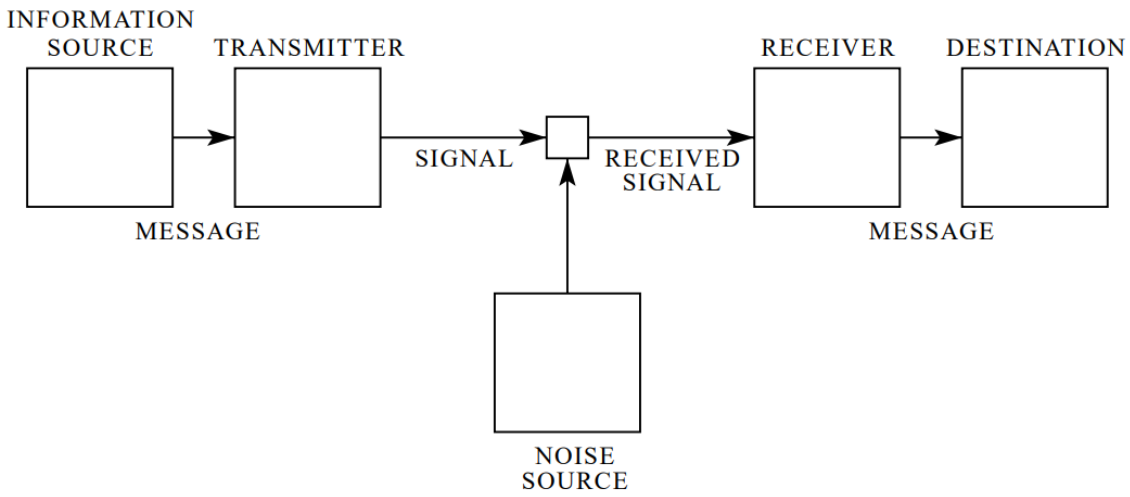


FIGURE 1.1. Structure of a communication system. Fig. 1 in [**Shannon and Weaver, 1948**].

Figure 1.1 shows the basic structure of a communication system as given in the paper.

The information source generates events in a probabilistic way, which produces a message that needs to be sent to the receiver. The transmitter encodes the message into a signal which is distorted by the noise source while being transmitted by the channel. Then receiver decodes the received signal and reproduces the original. Information theory is where the *bit* was invented, as a measure of information of a message. This means the information content of a message in bits is the number of simple chunks of binary information that needs to be sent in order for the receiver to reconstruct the message, following an **optimal** encoding-decoding schema for the problem. Hence we see an immediate mention of optimality as a way to solve arbitrariness in the fundamental step towards quantifying information.

The paper walks through -among other things- the process of finding an optimal encoding given the initial probabilistic information generation process. Notice that messages, signals or symbols used in the encoding are arbitrary, devoid of meaning. In fact, the main idea in quantifying the amount of information relies on the probability of a message occuring. The less probable an event is expected to occur, it is seen as having more information. Conversely, imagine we are encoding messages with a binary alphabet of 0 and 1. For more likely messages, we would want to use shorter encodings(e.g. '1' or '0' and not '110101' or '001010') because we expect to send them frequently, or with more probability. This implies, in information theoretic terms that, more probable messages contain less information than seldomly occuring messages. Again, the meaning of the message, the medium of transmission, the choice of encoding symbols/signals, physical instantiations of those signals, are all abstracted away. The famous Nyquist-Shannon Sampling Theorem even manages to get rid of problems (of communication theory) that are caused by the tension between discrete and continuous signals.

The logical structure that defined this process of communication is hence seen to be applicable in many domains, including activity of neurons, human memory, language acquisition,

genetics and the question of what it means for an image to be the image of a woman or a man, or of a dog or a cat. More on this in subsection 1.4.2

1.3 Cybernetics

1.3.1 Signals and Systems

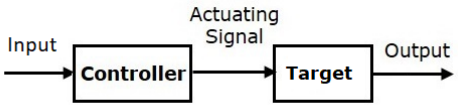
Signals and systems constitute the contemporary framework in which we situate the theoretical work originating in the Cybernetics community. The theory is currently referenced as dynamical systems theory, control theory and signal processing. The theory talks about signals: mathematical objects that are measurable, comparable, obeying certain logical axioms, with operations defined on them such as addition, differentiation, integration, time delay etc. Signals in this context, in essence are not different from the different kinds of messages/signals envisioned by information theory, which establishes the main connection between the two approaches. This means that cybernetical signals are also devoid of meaning and independent of physical instantiation or context.

Systems on the other hand, operate on signals. A system has input signals, which could determine the output in different ways, following rules with logical structures. Some concrete examples of systems could be an electronic circuit, with certain voltages or currents defined as input and output signal, and others as *internal* signals; or an aeroplane with the controls of the pilot defined as the input signals, and many other signals as output or internal signals. However, systems are strictly abstract mathematical objects, meaning that a system can be realized with a carefully arranged mass spring system, with correctly defined inputs and outputs; or an electrical circuit, without effecting any aspect of the system design, as long as the mathematical rules describing the relationships between the signals are similar. The transmitter, receiver and the sources in the basic information theoretic setup in Figure 1.1 are all also systems.

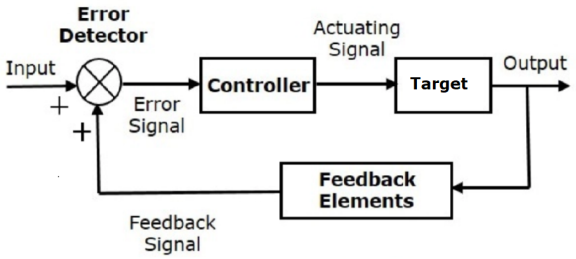
Given a target system, the task of control is to design a controller system that we plug to the target system such that we can easily drive the target system into the state that we want, or equivalently make it output the kind of signals that we want to get.

1.3.2 Feedback Systems

One of Cybernetics' most fundamental contributions was the realization of the effectiveness of feedback systems. In feedback (or closed loop) control systems, information about the current output signal is processed and supplied to the controller. This provides robustness to changes in the environment, modeling ambiguities, unmodeled disturbances and noise over signal transmission channels. This has pushed the cyberneticians to see their method as a way of general modeling of *adaptive*, *purposeful* or *intentional* behaviour. [**Dupuy, 2009**] Figure 1.2 shows schemas of open loop and closed loop control systems.



(A) Conventional schema of an open loop control system. Controller needs to be designed such that the input(desired behaviour) is transformed into an actuating (control) signal that drives the target system to realize the desired behaviour. This is sensitive to imperfect modeling, noise and perturbations.



(B) Conventional schema of a closed loop control/feedback system

FIGURE 1.2. Schemas of conventional control systems

1.3.3 Noise

In both frameworks, noise embodies the unmodeled, unpredictable or indeterminate parts of the task. In the mathematical domain, everything is well defined and most questions have clear answers, whereas applying this framework to real machines, and even to living beings as the title of Norbert Wiener's book *Cybernetics or Control and Communication in the Animal and the Machine*[**Wiener, 1961**] suggests, is another issue. Cyberneticians seems to admit the imperfections of their models, but suggest that it is robustness against noise that mostly defines what things are, potentially solving the missing meaning problem. We believe there is no better way to convey this than quoting Norbert Wiener's explanation:

> "The metaphor to which I devote this chapter is in which the organism is seen as message. Organism is opposed to chaos, to disintegration, to death, as message is to noise. To describe an organism, we do not try to specify each molecule in it, and catalogue it bit by bit, but rather to answer certain questions about it which reveal its pattern[1]: a pattern which is more significant and less probable as the organism becomes, so to speak, more fully an organism.
>
> We have already seen that certain organisms, such as man, tend for a time to maintain and often even to increase the level of their organization, as a local enclave in the general stream of increasing entropy, of increasing chaos and de-differentiation. Life is an island here and now in a dying world. The process by which we living beings resist the general stream of corruption and decay is known as homeostasis.
>
> ...
>
> It is the pattern maintained by this homeostasis, which is the touchstone of our personal identity. ... We are but whirlpools in a river of ever-flowing water. We are not stuff that abides, but patterns that perpetuate themselves.

[1]It must be mentioned that this explanation references the connections between statistical physics and information theory. Both theories have a notion of entropy, and statistical physical entropy can be seen as a concrete example of Shannon's entropy. The issue is very complicated and even famous polymaths like John Von Neumann are reported to admit that "nobody knows what entropy is"[**SVN,**]

> A pattern is a message, and may be transmitted as a message. How else do we employ our radio than to transmit patterns of sound, and our television set than to transmit patterns of light? It is amusing as well as instructive to consider what would happen if we were to transmit the whole pattern of the human body, of the human brain with its memories and cross connections, so that a hypothetical receiving instrument could re-embody these messages in appropriate matter, capable of continuing the processes already in the body and the mind, and of maintaining the integrity needed for this continuation by a process of homeostasis."[**Wiener, 1989b**]

As a consequence, with complex modeling, clever feedback mechanisms and better communication channels, any meaningful pattern can be captured and maintained by the cybernetic framework, and since it also captures some theories of physics, it is all too easy to reduce any concept, individual or entity to a set of patterns or to a set of yes-no questions about structure(giving 0 or 1 as answers, and thus constituting an encoding) that define it.

1.3.4 Communication and Control

Another thing that the title *Communication and Control in the Animal and the Machine* suggests is a connection between communication and control. We again leave the stage to Norbert Wiener's clear explanation:

> "When I communicate with another person, I impart a message to him, and when he communicates back with me he returns a related message which contains information primarily accessible to him and not to me. When I control the actions of another person, I communicate a message to him, and although this message is in the imperative mood, the technique of communication does not differ from that of a message of fact. Furthermore, if my control is to be effective I must take cognizance of any messages from him which may indicate that the order is understood and has been obeyed."[**Wiener, 1989a**]

1.4 Learning Machines

As digital computers became commonplace, and exponentially powerful with time, it became possible to process big, complex signals and to simulate complex systems. This practice culminated in the area of Artificial Intelligence. Machine Learning, in particular, is the area that is interested in general machines that learn from data to do some task: notice patterns, classify, fill in the blanks etc. Most of these methods have their basic background from signal processing and information theory. In fact, Cybernetics also generated the first notion of a mathematical neuron, a simple component that does a simple computation; and the notion of an artificial neural networks: systems constituted by connecting mathematical neurons in different configurations. [**McCulloch and Pitts, 1943**]

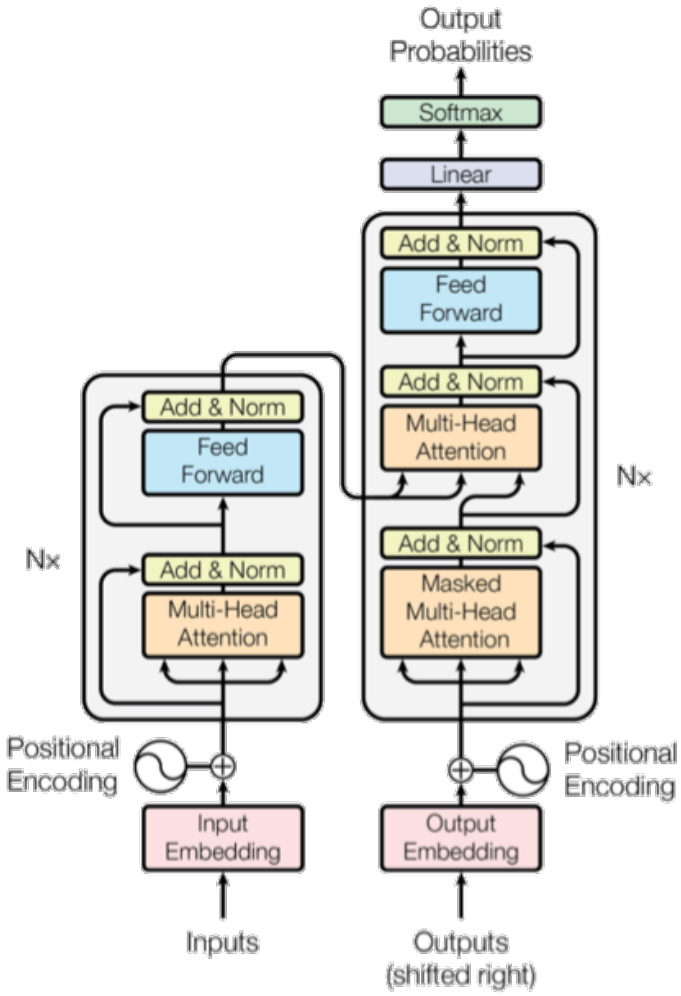


FIGURE 1.3. Basic Transformer architecture.[**Vaswani et al., 2017**] Notice the annotations "Nx" meaning the transformer cell shown in the figure is stacked multiple times to get a Transformer neural network. Each box represents an operation on the inputs, closely related to systems in the Cybernetics view.

Deep Learning, which is considered to be a subfield of Machine Learning, is interested in stacking many layers of neurons in a myriad of different ways, resulting in a deep neural network. Figure 1.3 shows the architecture of Transformer neural networks, the state-of-the-art method for time series and natural language modeling.

1.4.1 Just another box, but black

While deep learning fits well to the *boxes and arrows* models inherited from information theory and cybernetics, there is a shift in the modeling approach. While information carrying signals in Cybernetics are devoid of material dependency, in most of our applications, we know what they represent. They are mostly derived from some physical or mathematical

model relating to the physical instantiation of the system. However, with machine learning models, the system is realized as a computer program, as mathematical vectors composed of numbers, getting multiplied with matrices and going through arbitrary nonlinear functions depending on the choice of neuron models and network architectures. The signals flowing through the system are assumed to obtain meaning when the network is trained. First, the network architecture is determined. Afterwards, the training starts. That is, the parameters of the network are adjusted in an iterative manner, such that each iteration decreases the expected error of the decisions of the machine.

The error is computed with respect to a dataset that is used in training, most probably using some information theoretic error measure. In each training step, the processing done in the neural network changes, such that the final decisions of the network on the training data gets more and more correct (more and more correlated with the actual data, gains more and more *mutual information* with the dataset) with each adjustment. The mathematics of doing a single adjustment is basic[2], but why this simple approach works in insanely complicated tasks like face recognition or driving cars is a question that is wide open.

While our trust in our models in control theory generally come from consistent experimentation and equations of physics, validation of machine learning models is very problematic. The system is supposed to *generalize* the information it has learned while optimizing an error signal for the data it has seen. However, there is a risk that the system "memorises" the dataset, and doesn't really learn the underlying patterns. This is called *overfitting*. In practice, the basic process of validation is done by discarding a portion of the dataset, not showing them to the model during training. Then, different models are evaluated using these unseen datapoints to test how well they react to *new* datapoints.

However, this means we lose any and all abilities to interpret, interfere with or validate the actual flow and processing of information that is going on in the network, almost in all cases. We only have the intuitive understanding that the relevant information is distilled with each layer's computations and hopefully the output of the model captures the pattern we expect from the system.

Moreover, despite this lack of transparency or understanding, proponents of AI and automation argue that since the network learns directly from the input data, it is objective and free of subjective views that may be imposed on the machine in design-time. This, for example, manifests itself in the claim that "learning algorithms ... are not biased". [**LeCun, 2019**]

1.4.2 The paradox of designing an intelligent system

There seems to be another reason why validation of deep learning models is problematic. This stems from the range of problems we are applying them to. Success of these models has pushed engineers into applying these systems to tasks that do not clearly admit a well-defined ambiguity-free task definition in information theoretic terms. After all, when we have a clear

[2] https://www.youtube.com/playlist?list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi

FIGURE 1.4. Hand picked hard to recognize datapoints from the MNIST hand written digit recognition dataset.[**LeCun et al.,**][**Nielson,**] Can you recognize these digits? Can you blame the machines for failing? Can you praise them for not failing?

mathematical framework for a task, we find ways to either directly solve the problem, or gain considerable prior information about what the system should be doing. On another note: apart from some tasks that concern playing games where ability of agents are directly testable, it is not clear what it means for a machine to be better than humans at a task, however this is consistently reported.

This ambiguity is perhaps most obvious in the ambitious task of designing a thinking machine. While the well known Chinese Room Argument[**Cole, 2020**] argues that by its use of arbitrary and meaningless symbols, the information processing framework can not account for any understanding or thinking; Alan Turing suggests that a machine that is behaviorally indistinguishable from a person, by a person or a machine, can be said to be thinking, because it produces patterns that are indistinguishable from those of thinking beings. This has become to be known as the Turing Test. He even states: "The original question 'Can machines think?', I believe to be too meaningless to deserve discussion."[**Turing, 1950**] These two arguments seem to make very similar arguments about computing machinery: the Chinese room argument emphasizes that all machines can generate are patterns, Turing emphasizes that machines can generate all patterns. They conclude in two completely opposite positions.

A useful endeavor, we find, is to ask "Can machines achieve superhuman or superbiological performance in the task of dying[3]? And if not, does that mean death is meaningless?"

Still, even in the fairly simple task of recognizing hand written digits, there are ambiguous edge cases. Figure 1.4 shows some examples. Neural networks are today applied to generate text descriptions of images, or to evaluate how fit people are for professional positions. These tasks are clearly not easily expressible in information theoretic terms. This results in impossibility of the evaluation of the model's quality.

Actually, if one assumes that any relevant aspect of any phenomenon can be captured by the syntactical information framework, then a neural network that classifies hand written digits can be thought of as learning the patterns that define what it means to be a hand written digit 8, in connection to Wiener's argument presented in subsection 1.3.3. This would be cutting off all references to context, nuance and meaning, allocating all that into the dataset's structure: the mathematical relationships between the datapoints and patterns present in the dataset.

[3]The grim tone is not intentional, but necessary because the task of living is very closely related to the information processing capacity of the agent in its cybernetic definition.

This ambiguity also presents itself when we think about *overfitting*. The machine is expected to learn patterns from the training data that are present in the rest of the data in the world, and not fixate on the patterns that appear only on this subset of the dataset, meaning not fixate on the noise introduced by random sampling of the input data.

CHAPTER 2

Urgency: Political Implications of Intelligent Systems

2.1 Meaning Making

The first advocates of automation were probably thinking about abolishing heavy human labor, or waste of human time by making machines autonomously take care of trivial, redundant and/or heavy work. With the advancements in computing technologies, it became possible to assign tasks that require human cognitive capacities to machines. Moreover, as argued in chapter 1, with systems that learn from real world data, an argument for the objectivity of these systems' decisions became possible. After all, the data is coming from the real world, and the system designer doesn't have a say in what is learned and what is not.

A political critique of this argument can be:

> "The paradox of Big Data is that it both affirms and denies this 'interpretative nature of knowledge.' Just like the Oracle of Delphi, it is dependent on interpretation. But unlike the oracle priests, its interpretative capability is limited by algorithmics—so that the limitations of the tool (and, ultimately, of using mathematics to process meaning) end up defining the limits of interpretation. Similarly to Habermas, Drucker sees the danger of 'ceding the territory of interpretation to the ruling authority of certainty established on the false claims of observer-independent objectivity.'"[**Cramer, 2018**]

Recall also that we don't have a real explanation of how complex machine learning systems function, when they could fail and how exactly they fail. This is actually part of the reason the system designer can not interfere with the system, thus making the system *objective*. However, in this case it is impossible to attribute any responsibility to any party that is actually able to respond. This fact, combined with the illusion of objectivity, make a very dangerous combination.

We are invited to take a positive stance to the question of availability of objective information, that accepts no questioning or rationalization, other than the cybernetic meaning-free syntactical worldview of information bearing signals and information processing systems. When taken to its extreme, this suggests that machines can surpass human understanding in any context, since meaning is defined by regularities that persist in the data that is generated in the world. In general, an emerging urgency concerning applications of intelligent systems to real-life decisions seems to show itself in this context.

Urgency: Human beings are being replaced by unexplainable, unaccountable machines

in their role as message/signal/data interpreters, in their responsability to make[1] meaning out of shallow information.

We now try to give several examples of how this can lead to political issues.

2.2 Dirty Data

Potentially the biggest and most obvious problem with the lack of understanding in an artificially intelligent system is it's ignorance of the real world's pathetic state and what the *right decision* is.

Automatically gathered data can have many problems.[**Steyerl, 2018**] First of all, it inherits any commonplace biases that the environment it is generated in has. Examples could include: black people being flagged as potential criminals more often than white people, despite lack of difference in behaviour; women being consistently classified as lower quality workers compared to men, despite lack of difference in qualifications.[**Cramer, 2018**] This is of course, just another cause of noise.

Since noise needs to be eliminated from a dataset, any real outliers are also possibly interpreted as dirty datapoints[2] and the model is encouraged to ignore them for the sake of overall accuracy optimization.

Proponents of automation and strong AI insist that an abundance of clean data will solve all problems, because it will diminish *noise* by introducing *redundant information* and hence decreasing *entropy* of the dataset, making it easier to *predict*. It must be noted that this argument only restricts the notion of noise. It assumes that an abundance of clean data will include any and all variations and patterns that are acceptable. This means any true outlier, that is not part of an abundance of similar datapoints, does not deserve recognition. This is plainly false in many areas where the world is envisioning applying these systems to. It is true in a highly limited set of contexts.

Finally, the concept of clean data is shady in itself. Data always needs to be cleaned. If it is cleaned by humans, then an obvious weak point for the objectivity argument is introduced, and the subject becomes immediately political. If it is cleaned by machines that are trained on other data, then the problem of dirty data is not actually solved.

2.3 Automated Decision Making

Automated Decision Making is the most obvious application where the lack of a responsible party causes problems. Most of the time, these applications also suffer from the lack of clarity in the task definition mentioned in chapter 1. Take for example students that are

[1] Meaning making is originally a term that is used in semiotics: "the study of signs and sign-using behaviour",[**Brittanica, 2020**] and sign processes as the origin of meaning. Semiotics proposes a completely different reduction and methodology which is certainly relevant, but unfortunately falls out of the scope of this text and our knowledge.

[2] Glitches? https://www.youtube.com/watch?v=DqNPgd5B3io&t=424s

subject to automated grading.[**Mahdawi, 2020**] There is no real way to justify the grades to the students, or to answer questions they might have. In addition, since we don't already have an unambiguous or objective way of defining criteria for which students should get what grade, it is impossible to validate the quality of the decisions taken by the machine in a decisive fashion, that doesn't take more time and energy than actually grading the students by hand.

2.4 Gamification - Machinization

The cybernetic view that things can be abstracted and reduced to information by measurements, and that this information can be processed and organized to generate an efficient purposeful system, also suggests organizing human behaviour around measurements, signals and predictions. An instance of this is gamification, under which agents are put in a game-like context in which they behave to win, improving some overall objective like throughput or efficiency.[**Han, 2017**]

For example, Amazon uses a scoring system on its employees that determines the continuation of their employment.[**Samuel,**] Indeed, this is obviously cruel, and many performance measures basically draw on the same idea. Grading systems in schools, academic scoring systems, insurance scores all draw on the idea that the relevant information about an individual can be measured reliably from their output.

While this may be helpful in certain specific senses, it encourages people to measure their own output's value in terms of the syntactic gamification system they are playing in. An example of this is the abundance of low quality papers that repeat themselves, or the phenomenon that many papers start resembling each other, in some academic areas: what could be speculatively described as *intentional* overfitting.

This political aspect of the Cybernetics view does not necessarily relate to artificial intelligence systems, however it is plausible and expected that this setup pushes participating agents to behave the same way an artifically intelligent agent is striving to behave: minimizing error or maximizing return in a (hopefully) *well defined* input-output relationship, resulting in a machinization of living beings, and hence completing the loop to render the information theoretic account of human cognition and behaviour a self-fulfilling reduction.[**Denizhan, 2014**]

CHAPTER 3

Intervention: Cancelling Noise Cancelling

As a personal intervention, I activated the *Ambient Sound* mode on my headphones, which turns off noise cancelling, and amplifies the environmental sounds, adding them to the music you are listening. The motivation for this mode is to be aware of surrounding sounds, like vehicles or people talking to you. Contrary to almost all applications of signal processing or machine learning, this application concretely has the signal and noise seperated. The signal is already digitally encoded and seperately supplied via the headphone cable, while the noise can be pretty accurately measured via a microphone. We know the listener is hearing some mixture of the two. Thus the task definition of noise cancelling is pretty clear and the task of adding the noise to the signal is very easy. In this regard, it can be said that I could see how I react to more noise, without much ambiguity. However, one moment where I mistook the noise coming from a street musician as part of the signal coming from Spotify suggests that, even though I *know* the definitions of supposed signal and noise, I lack the ability to perfectly classify input as noise and signal.

I normally use my headphones a lot: whenever I study or am walking/commuting alone. I have come to appreciate the value of noise cancelling technologies in the bus or the U-bahn. Apparently, it is quite frequently impossible to listen to music privately without some blocking of background noise in the modern city.

However, this was much less noticeable while studying in crowded spaces. Focusing on coursework seems to activate a natural noise cancelling, which suggests yet another way to define noise and signal, noise and signal as seperated by the user's cognition. This would mean that the Spotify signal is noise whenever I focus on the subject I am studying, and becames the primary signal when I take a break, tying the definition deeply to my actual state of mind and the context in which we are using the terms. It is also instructive to think about the differences between this kind of noise cancelling, and a literal cancelling of the noise at the input level.

CHAPTER 4

Design Project: FaceShift

4.1 Motivation

The basic model of communication from information theory, the cybernetic outlook and artificial intelligence models are all captured via the same imagery of boxes and arrows, because they define two distinct entities that govern all meaning in the world: numerical/logical/syntactical signals and processors that transform these signals into new signals, changing the information content or encoding. In fact, a dependence on purely formal relationships and structures are visible from the etymology of the words system[1] and information[2].

This raises two potentially promising directions to explore.

4.1.1 Exploring clashing semantic spaces

An immediate urge stems from the observation that the information theoretic narrative rests on a mutual agreement of encoding and decoding schemes. Even in deep learning systems where signals are farthest from being interpretable, we have the understanding that the neural network components learn how to communicate signals to each other in a way that makes internal sense. However, mixing signals from two independent learning systems don't need to be encoded in the same way, or interpretable in the same fashion. Thus, an interesting experiment would be to materialize the outcome of neural networks interpreting signals that come from another system, another encoding-decoding scheme, another meaning space.

4.1.2 Purposefulness/Intentionality Recognition

An interesting question is: can living beings deduce the meaning of the computation being done in complex, opaque learning systems via experience, interaction or cooperation?

This could potentially be investigated by situating the experimenting natural agent as the input signal generator. Supplying the decisions (outputs) of the system to the experimenter, we can hope for an understanding of the process, explanations of the decisions or a refutation of the capacity of purposeful/meaningful behaviour of the machine to emerge.

[1] "**system(n.):** ... from Greek *systema* 'whole, a whole compounded of parts,' from stem of *synistanai* 'to place together, organize, form in order' "[**Harper, b**]

[2] "**information(n.)** : ... noun of action from past participle stem of informare 'to train, instruct, educate; shape, give form to'"[**Harper, a**]

Current efforts in understanding or explaining deep learning models constitute the field of *Explainable AI*. Explainable AI faces even bigger task definition crises than AI: how can one test if a machine has learned the complicated task with the ambiguous task definition, how to evaluate the qualities of any explanations of the decisions of the networks, what are the connections of the explanations to the abstract mathematical process that governs training?

Being recently established, it lacks sophisticated frameworks, and mostly consists of independent methods being designed by independent groups. It can be said that the explanation methods envisioned are all mathematical in nature, with references to information, signal processing and optimization theoretic terms. Hence, methods are designed to be applicable to general classes of models, data and tasks.

What we mean by purposefullness/intentionality recognition is the opposite stance. The idea is to try to provide intuitive explanations based on personal experience, to specific singular systems that we have in front of us, imitating an anthropologist trying to understand a new mode of being, or a clinical psychologist trying to come to terms with their patient.

4.2 Technical Background

4.2.1 Dimensionality Reduction

Dimensionality reduction is a machine learning task. In most basic terms, it can be defined as the task of data compression.

Consider a color image that is 100 pixels in both height and width. This means there are $100 \cdot 100 \cdot 3 = 3 \cdot 10^4$ degrees of freedom, 30000 numbers that precisely define the input image in the *set of all possible messages*, or in contemporary lingo, in the input space. The information in this case is distributed in many pixels, and we are looking for a better, compact encoding such that we can *reconstruct either perfectly or approximately the input image upon receiving the compact encoding.*

4.2.2 Data Spaces

This narrative of having better and better encodings that carry more and more information and less noise about the correct answer is the main intuitive explanation we have for how deep learning models manage to solve problems. However we lack the complementary theoretical framework and definitions to state the same thing in a mathematical language. Information theory sets clear limits to the problem of compression. Still, learning machines don't need to expect any input whatsoever. A face recognition system is expected to work on pictures of faces, which means defining the set of possible inputs as "all images" is unneccessary. With dimensionality reduction, given a dataset, we are interested in finding short encodings that capture all the relevant information for that dataset, capturing the relations between the datapoints and any structrue that may be present in the dataset, any patterns that might be useful for reconstruction.

Dimensionality reduction algorithms consist of two components: encoders, which take in raw data and spit out compact encodings, and decoders which take encodings and successfully reconstructs the input features. Note that the encoding need not have any connection to a material reality, a decoder can decode any encoding into a hopefully meaningful output signal.

4.2.2.1 Principal Component Analysis

Principal component analysis(PCA) is the simplest method for dimensionality reduction. It is also the case of an interpretable model: where we have a theoretical understanding of the process, we know what we are optimizing and we can find unique optimal solutions to problems. In theoretical terms, PCA finds the best linear subspace such that projections on this space results in minimal error.

In order to demonstrate intuitively what PCA does, Figure 4.1 shows two cases of applying PCA to reduce a dataset from two dimensions to a one dimensional representation.

4.2.2.2 Autoencoders

Autoencoders, quite simply, use deep neural networks as encoders and decoders, and optimizes for reconstruction error. Figure 4.2 shows a basic diagram. The icons are taken from [**Rocca, 2019**] and will be used throughout the chapter.

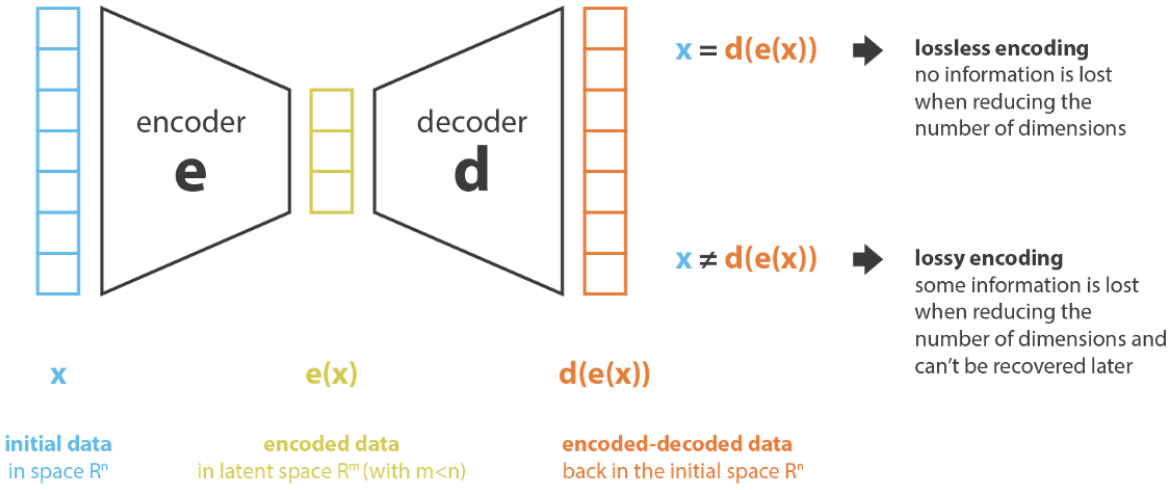


FIGURE 4.2. Boxy representation of autoencoders from [**Rocca, 2019**]

4.3 Design

4.3.1 Autoencoders

FaceShift uses two autoencoders trained on two different tasks. We feed the encodings coming from one encoder-decoder couple to the decoder of another. This corresponds to defining a semantic shift, interpretation of a signal in another domain with another meaning or

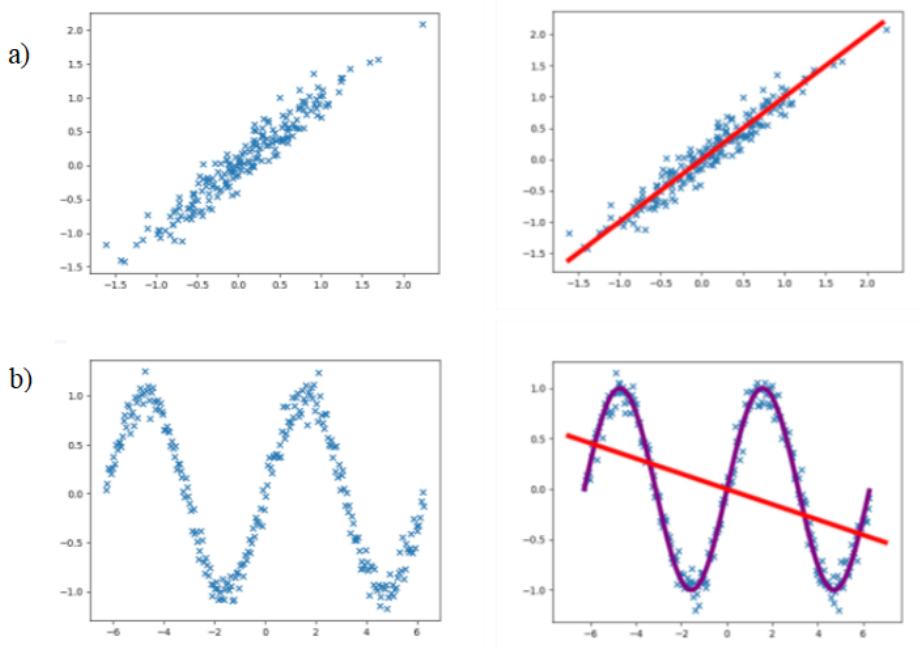


FIGURE 4.1. Example applications of dimensionality reduction. Each $\times$ is a datapoint.
a) PCA can find the best straight line that fits the data. The red line is the new one dimensional space to represent the dataset in. If we map each data point to the closest point on the red line, we can identify each datapoint with a number depending on its location on the line. This preserves much of the information in the dataset: relationships between datapoints and patterns that may be present in it. Decoding process is to take the one dimensional encoding (location on the line) and to generate the same position in the original 2D coordinate system.

b) PCA can only find straight surfaces, which is why we have complete theoretical understanding of its functioning. More complex topologies necessitate more complex models. Neural networks have a wide range of expression in this regard, but the price is that we don't know what the resulting dimensionality reduction means, other than that it minimizes our error measure.

another physical instantiation. If the transfer of encodings from one autoencoder to the other is also done by a dimensionality reduction algorithms, all steps of the FaceShift architecture have some justification for meaning preservation in information theoretic terms.

4.3.1.1 Face Autoencoder

The first autoencoder we use encodes low resolution pictures of human faces to a 100 dimensional encoding space. A pretrained neural network trained on a celebrity face dataset.

We have access to both the encoder and the decoder. The encoding is hoped to be representative of the input face features, as well as details like orientation, lighting, position.

4.3.1.2 Sound Autoencoder

We use the architecture described in [**Engel et al., 2020**]. This work is very interesting that it encourages combining classical, interpretable signal processing techniques(synthesizer and filters) in combination with deep learning models, assigning the "hard, ambiguous" part of the task description to the neural network while having some interpretation of what the system is doing.

Briefly, the authors use the autoencoder framework to design a machine that learns to generate any melody in a specific instrument's sound.

This is accomplished by first recording a dataset of a single instrument's solo recordings. Then, a conventional signal processing system(that doesn't rely on data or learning) finds numerical values standing for the pitch and volume at each point in time. *These two values are the encoding of the input recording.* Afterwards, a neural network processes the encodings to generate sound.

This means that in this design, the encodings are actually meaningful. Each coordinate of the 2D embedding has a physical meaning that will be present in the output sound, even though we are oblivious to how the sound is generated from these encodings.

While the model is trained with recordings of a single instrument, since we have a clear explanation of the encodings, we can use any input sound to find the pitch and volume. We can then generate a violin, trumpet, flute or a saxophone that plays the sound.[3]

4.3.2 FaceShift

FaceShift is a computer program that attempts to create a space of speculation over semantics of signals. It uses an autoencoder trained to encode faces to obtain a 100 dimensional embedding of your face. Then, encoding schemes learned by PCA are applied to project this embedding on a 2D subspace, further reducing dimensionality. 3 representations of *your face* are visible on the screen:

(1) **Webcam Video:** A rectangular frame from your webcam is cropped. The intended use is to put your face in this area.

You can enlarge the rectangle of the frame relative to the whole webcam input by pressing "X" and shrink it by pressing "Z". You can use the "W, A, S, D" keys to move the frame's location on the webcam video. Only the cropped frame is shown

[3] https://magenta.tensorflow.org/ddsp has some cool examples. https://colab.research.google.com/github/magenta/ddsp/blob/master/ddsp/colab/demos/timbre_transfer.ipynb lets you easily generate sounds from your recordings. The inputs sounds should be monophonic.

in FaceShift, but using these keys, it is easy to see your face on the screen in the scale that you want.

(2) **Reconstruction:** Your face is encoded and decoded by the face autoencoder from subsubsection 4.3.1.1. The decoded version, or the reconstruction of your face is shown next to the webcam input.

(3) **2D Plot:** A 2 dimensional coordinate system is also displayed. A dot shows the encoding of the video input, reduced to 2 dimensions, by default using the PCA component. There is also the option to experiment with a random dimensionality reduction. A red point indicates that PCA is being used. A yellow point indicates a random transformation is used. You can press the *space* button to flip this setting to experiment with the effectiveness and meaning of PCA versus a random transformation.

The origin of this plot(the zero encoding) corresponds to a pitch of 440 Hz and a volume value of 0.5 in range $[0, 1]$.
The coordinates are also scaled, and clipped to be in the required range. Pressing "V" scales up the coordinate values while pressing "C" scales them down.

When you press 'R', a blue dot appears on the top left to indicate FaceShift is recording.

Your movements are recorded and when you press 'R' again, the recording stops and FaceShift starts decoding the embeddings for each frame. At the and, a video with the generated sound is displayed and also saved in the file *output.mp4*. Graphs of the recorded pitch and volume signals are also saved in *pitch.png* and *volume.png*

In order to download, run and use FaceShift, please read the instructions here[4].

Figure 4.3 shows the system schema that describes FaceShift.

4.4 Resources and References

The original codebase for the face autoencoder is from this[5] GitHub repository. The model is trained on the CelebA dataset [6].

Here[7] is a quick demo of an application of the instrument autoencoder, and here[8] is the full Differentiable Digital Signal Processing library.[**Engel et al., 2020**]

The libraries Pytorch and Tensorflow are used in these implementations respectively. The sklearn package was used to compute the PCA transformation from the encodings of the CelebA dataset.

[4] https://github.com/gumityolcu/FaceShift
[5] https://github.com/matthew-liu/beta-vae
[6] http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html
[7] https://colab.research.google.com/github/magenta/ddsp/blob/master/ddsp/colab/demos/timbre_transfer.ipynb
[8] https://github.com/magenta/ddsp

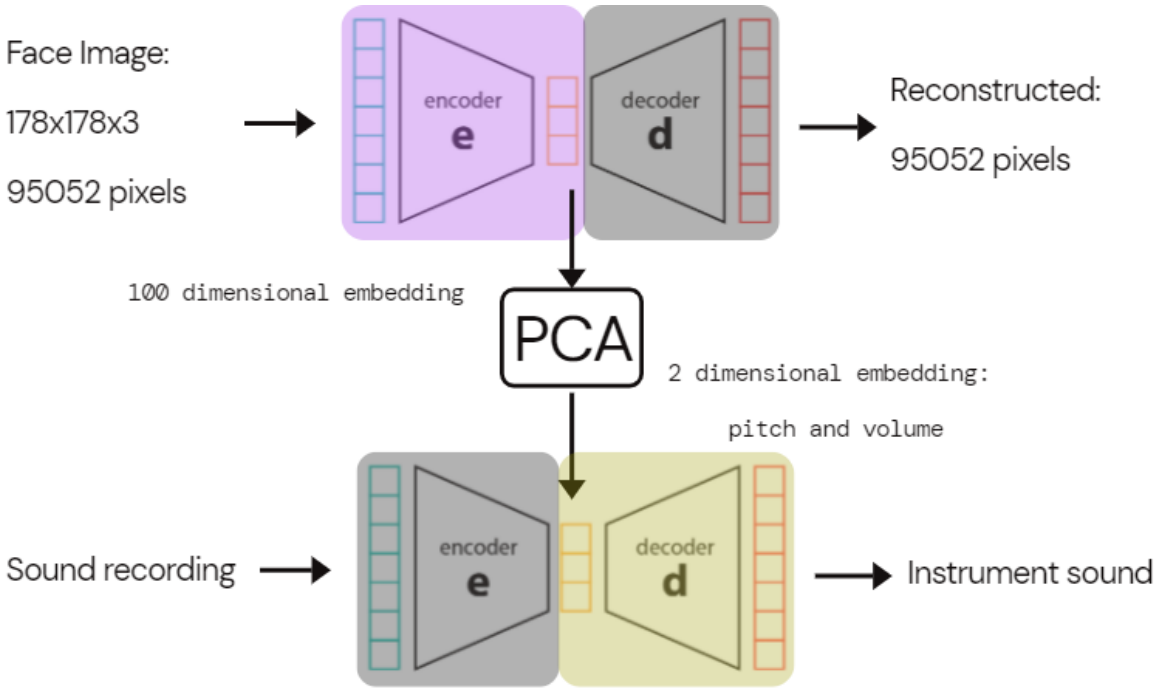


FIGURE 4.3. System schema of FaceShift.

CHAPTER 5

Reflection and Conclusion

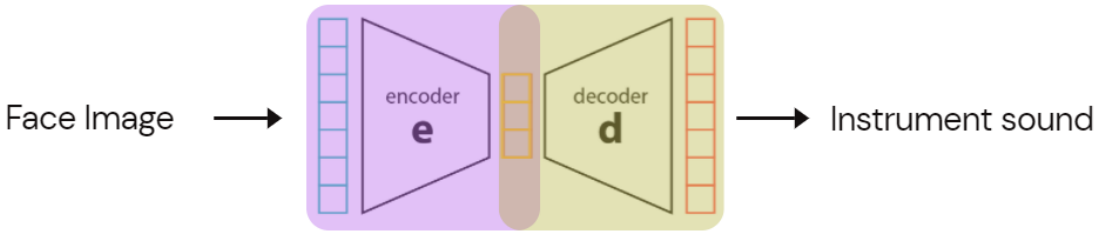


FIGURE 5.1. Simpler schema of FaceShift

FaceShift bundles four representations of the same information. It attempts to create a space to speculate about what signals may mean when interpreted in different learning contexts.

(1) **Webcam Video:**
The first representation can be thought of as the input: it is what we have control over. Then, this input is encoded by the face encoder, and all other representations are actually generated from this information. Thus, FaceShift can be seen as presenting different manifestations of what information the encoder detects in your face. Then, these representations could be used to understand, explain, get-to-know them, assign meanings to certain aspects of the representations.

This argument would be much more stronger with a fully-real time system that also generates real time sound. The current flow of "start recording, make move, stop recording, process, analyze" is not reflective of the intention of involving intuitive and potentially primal abilities into play. Real time audio generation is complicated to program, and may be impossible with complex deep networks, without access to sophisticated machines.

(2) **Image Reconstruction:**
In general, the reconstructed images seem to be very poor and some undesirable symptoms are present. If your face is bigger than expected it is interpreted as the background, and it can actually reconstruct detailed backgrounds in surprising detail, compared to actual facial features. It is also extremely dependent on lighting.

However, poor performance of the face autoencoder could also be an advantage for the previous point. Some obvious pitfalls or overly basic strategies that the encoder may employ might provide an easier environment to make the point of interacting with the machine to understand its *reasoning*.

For example, one pattern we recognize just by interacting with the reconstructed image is that, generally the orientation of the reconstructed face does not depend on the input face's orientation, but is conditioned on the position of the face. If you move left, without turning your head, the reconstructed image stays in the middle but the reconstructed face is rotated.

The decoder always puts the face in the middle in the reconstructed image, whatever the input.

(3) **2D Plot:**

The 2D plot is the most abstract of all representations. However it is the input to the decoder, which assigns it a meaning as a way of visualising your input to the sound generation system: the input produced by your face, the encoder, and the PCA process. This plot is also useful to quickly show that PCA does capture *some* signal. You can press the *space* key to change to a random dimensionality reduction instead of PCA. In practice, the noise from the camera seems to affect the 2D representation much more, and moving your head around changes the 2D encoding drastically, seemingly randomly. PCA's behaviour is stable. Camera noise has virtually no effect. Moreover, you can explore a much smaller portion of the encoding space. The machines seems to reliably assign each face a relatively small area in the 2D embedding space, which may be tempting to be interpreted as meaning(in fact, this directly implies high information in the information theoretic framework, because *being in a smaller region of the space* may be interpreted as *having a small probability and thus high information value* under basic assumptions).

(4) **Sound Reconstruction:**

The sound reconstruction decoder has a very clear input definition, however the output should be an instrument(let's assume trumpet) playing the given pitches with given amplitudes. While some instruments can only generate certain pitches, all performances used in training the sound autodecoder use the Western 12-tone-equal-temperament note and tuning system. Our encodings, however, are intended to be continuous. In practice, they span all pitches in a certain range, up to a certain precision which is much bigger than 12-per-octave.

The decoder is expected, then, to *trumpetize* the given encodings as much as possible and as much as it has learned to, which may end up sounding like a broken violin.

This suggests that FaceShift can be seen as an instrument. You are constrained/encoded to a portion of the space, which you have some control over by scaling. Then your movements are translated to sound. It is up to you to make it sound like a trumpet or a broken violin (whichever you want) by behaving the required way. The extent to which your intention is realized depends on how well you communicate with the encoder to have control over the 2D representation; and then with the decoder in order to actually produce the music you want to produce. Again, FaceShift, also determines the music by constraining your space of possibilities. This, combined with the reflections in (1), could give a use-case to having an intuitive understanding of a singular neural network's operation. This argument would also be much easier to make if the system was working in real time.

As explained in (3) and (4), an interaction with FaceShift is personal in some sense. Also, the limitation of the encoding limits the output space. This urges us to think about multiagent applications. 5 people's faces could be mapped to 5 different instruments to generate a soundscape of the group, with minor control given.

Another limitation of FaceShift's ability to push further on the issue of meaning is the decoder. The current open-source decoder is trained on a single instrument. If it were trained with many instruments, using more than 2 dimensional encodings then we could use all the information in the encoder's output, not just a 2 dimensional representation. Furthermore, the decoder wouldn't be trying to produce the sound of a single instrument, but would be trained to handle many different encodings of many instruments. Thus, more overtly ambiguous regions of the encoding space would be present, and this would more closely correspond to *the meaning of a face image as an instrument sound, defined by the encoding-decoding processes of certain neural networks*.

Training this system is easy with access to a high end GPU. It should be possible on a small scale using free cloud computing services. We will be trying out simple versions of this design.

Speculative Concepts

- **Knowable/Understandable AI:** As described in subsection 4.1.2, it seems promising to use intuition and physical experience-history as a conscious being in building trust towards the decisions of a system, or at least understanding or empathizing with them. This could be described as putting *getting-to-know* or *empathizing* over logical explanation. The endeavor could be called Knowable or Understandable AI as opposed to Explainable or Interpretable AI.

- **Intentional Overfitting:** Machine learning engineers do sometimes intentionally design systems that overfit, for example to get a model that has learned to recognize one object very well. However, intentional overfitting as a concept is not present in a mathematical sense. While we think of it to basically mean the "self-fulfilling reduction" of [**Denizhan, 2014**], it is a nice wording in that it points out the absurdity in using straightforward accuracy measures while expecting the model not to overfit[1]. In general, it may be useful to argue that learning and overfitting are not conceptually clearly seperated.

- **Morphological Turn:** Similar to the linguistic turn that took place in the early 20^{th} century, the dominance of cybernetic conceptions in science and philosophy of the late 20^{th} may be described as a *morphological turn*: characterized by a focus on the formal aspects of processes in the world as a primary model.

- **Self-cancelling/self-invalidating Reduction:** We already talked about self-fulfilling reductions, similar to self-fulfilling prophecies. The idea of a counterpart to self-cancelling prophecies is interesting. It seems to make political sense to think about the possibility of turning the morphological turn into a self-cancelling reduction.

[1] In fact, the current central question of statistical learning theory is why overfitting does not happen with some of the very complex models like deep neural networks: https://arxiv.org/abs/2112.03968

Course Evaluation

The course was enriching. I gained experience in writing, explaining, organizing ideas, understanding/formulating political arguments, interacting with political texts, understanding and critiquing designs. In general, the freedom and the encouragement you provide is also super-nice.

I struggled with the practical aspects of the course. I struggled with coming up with an intervention, and reflecting on it. It was not exactly because I didn't understand the task. I think I understood it as clearly as it was defined. I was focusing on the concept of noise and it was very hard to find an interesting intervention or experiment that related, and even harder to reflect on my experience. The design process was also tough. It was constrained heavily by temporal and practical constraints. And in the end, I turned back to the technical arsenal I had from my previous knowledge about philosophical repercussions of general artificial intelligence.

I am conflicted about how this turned out. On the one hand, I would have liked to avoid arguing against a single group of western scientists, and the whole narrative could be not constructed around their terminology. It could also have saved 10 pages from this text. On the other hand, I have spent the last two years studying dynamical systems, control theory, information theory and deep learning. I am also aspiring to work in explainable AI. One of my motivations in taking the course was to gain new perspectives on the stuff I was studying. This course provided a platform to do exactly this, and to formulate the problems I have with the concept of information in the form of a written text and a design project. I could not have done that without the structure, guidance and interaction provided by the course. And the process, although I say was tough, was much more fluid and fun than I would have expected. So I am also very happy with what I spent time working on. Frankly, I also really don't know what I would have talked about concerning politics of technology and body or space.

References

[SVN,] Neumann-shannon anecdote. http://www.eoht.info/page/Neumann-Shannon%20anecdote. Accessed: 04.03.2022.

[Brittanica, 2020] Brittanica, E. (2020). semiotics. https://www.britannica.com/science/semiotics. Accessed: 04.03.2022.

[Cole, 2020] Cole, D. (2020). The Chinese Room Argument. https://plato.stanford.edu/archives/win2020/entries/chinese-room/.

[Cramer, 2018] Cramer, F. (2018). Crapularity hermeneutics: Interpretation as the blind spot of analytics, artificial intelligence, and other algorithmic producers of the postapocalyptic present. In *Pattern Discrimination*, chapter 2, pages 23–58. meson Press.

[Denizhan, 2014] Denizhan, Y. (2014). Performance-based control of learning agents and self-fulfilling reductionism. *Systema: connecting matter, life, culture and technology*, 2(2):61–70.

[Dupuy, 2009] Dupuy, J.-P. (2009). A poorly loved parent. In *On the Origins of Cognitive Science: Mechanization of the Mind*, chapter 2, pages 43–65. MIT Press, Cambridge, MA.

[Engel et al., 2020] Engel, J., Hantrakul, L. H., Gu, C., and Roberts, A. (2020). Ddsp: Differentiable digital signal processing. In *International Conference on Learning Representations*.

[Han, 2017] Han, B.-C. (2017). Gamification. In *Psychopolitics: Neoliberalism and New Technologies of Power*, chapter 10, pages 49–54. Verso books, London.

[Harper, a] Harper, D. Etymology of information. https://www.etymonline.com/word/information. Accessed: 05.03.2022.

[Harper, b] Harper, D. Etymology of system. https://www.etymonline.com/word/system. Accessed: 05.03.2022.

[Heims, 1989] Heims, S. J. (1989). Introduction. In *The Human Use of Human Beings: Cybernetics and Society*, pages xi–xxv. Free Association Books, London.

[LeCun, 2019] LeCun, Y. (2019). Tweet on 07.12.2019. https://twitter.com/ylecun/status/1203211859366576128. Accessed: 04.03.2022.

[LeCun et al.,] LeCun, Y., Cortes, C., and Burges, C. J. C. Mnist handwritten digit database. http://yann.lecun.com/exdb/mnist/. Accessed: 04.03.2022.

[Mahdawi, 2020] Mahdawi, A. (2020). It's not just a-levels – algorithms have a nightmarish new power over our lives. https://www.theguardian.com/commentisfree/2020/aug/19/its-not-just-a-levels-algorithms-have-a-nightmarish-new-power-over-our Accessed: 04.03.2022.

[McCulloch and Pitts, 1943] McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

[Nielson,] Nielson, M. Deep learning. http://neuralnetworksanddeeplearning.com/chap6.html. Accessed: 04.03.2022.

[Peters, 2012] Peters, B. (2012). Normalizing soviet cybernetics. *Information & Culture*, 47(2):145–175. Excerpt from Rosenthal, Mark M.; Iudin, Pavel F., eds. (1954). Kratkii filosofskii slovar [Short Philosophical Dictionary] (4th ed.). Moscow: Gospolitizdat. pp. 236–237.

[Rocca, 2019] Rocca, J. (2019). Understanding variational autoencoders (vaes). https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73.

[Samuel,] Samuel. Amazon point system- how and when does amazon assign points? https://howigotjob.com/uncategorized/amazon-point-system-how-and-when-does-amazon-assign-points/. Accessed: 04.03.2022.

[Shannon, 1956] Shannon, C. E. (1956). The bandwagon. *IRE transactions on Information Theory*, 2(1):3.

[Shannon and Weaver, 1948] Shannon, C. E. and Weaver, W. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

[Simon, 2003] Simon, H. A. (2003). Information processing. In *Encyclopedia of Computer Science*, page 856–858. John Wiley and Sons Ltd., GBR.

[Slamecka, 2018] Slamecka, V. (2018). information processing. https://www.britannica.com/technology/information-processing. Accessed: 03.03.2022.

[Steyerl, 2018] Steyerl, H. (2018). A sea of data: Pattern recognition and corporate animism (forked version). In *Pattern Discrimination*, chapter 1, pages 1–22. meson Press.

[Turing, 1950] Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236):433.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

[Wiener, 1961] Wiener, N. (1961). Cybernetics: Control and communication in the animal and the machine.

[Wiener, 1989a] Wiener, N. (1989a). Cybernetics in history. In *The Human Use of Human Beings: Cybernetics and Society*, pages 15–28. Free Association Books, London.

[Wiener, 1989b] Wiener, N. (1989b). Organization as the message. In *The Human Use of Human Beings: Cybernetics and Society*, pages 95–105. Free Association Books, London.