Chapter 1

Concepts of Information

1.1 How to talk about information: Some simple ways

The epigraph to this Part is drawn from Strawson's contribution to his famous 1950 symposium with Austin on truth. Austin's point of departure in that symposium provides also a suitable point of departure for us, concerned as we are with information.

Austin's aim was to de-mystify the concept of truth, and make it amenable to discussion, by pointing to the fact that 'truth' is an abstract noun. So too is 'information'. This fact will be of recurrent interest in the first part of this thesis.

" 'What is truth?' said jesting Pilate, and would not stay for an answer." Said Austin: "Pilate was in advance of his time."

As with truth, so with[1] information:

> For 'truth' `['information']` itself is an abstract noun, a camel, that is of a logical construction, which cannot get past the eye even of a grammarian.
>
> We approach it cap and categories in hand: we ask ourselves whether Truth `[Information]` is a substance (the Truth `[the information]`, the Body of Knowledge), or a quality (something like the colour red, inhering in truths `[in messages]`), or a relation ('correspondence' `['correlation']`).
>
> But philosophers should take something more nearly their own size to strain at. What needs discussing rather is the use, or certain uses, of the word 'true' `['inform']`. (Austin, 1950, p.149)

A characteristic feature of abstract nouns is that they do not serve to denote kinds of entities having a location in space and time. An abstract noun may be either a count

[1] Due apologies to Austin.

noun (a noun which may combine with the indefinite article and form a plural) or a mass noun (one which may not). 'Information' is an abstract mass noun, so may usefully be contrasted with a *concrete* mass noun such as 'water'; and with an abstract *count* noun such as 'number'[2]. Very often, abstract nouns arise as nominalizations of various adjectival or verbal forms, for reasons of grammatical convenience. Accordingly, their function may be explained in terms of the conceptually simpler adjectives or verbs from which they derive; thus Austin leads us from the substantive 'truth' to the adjective 'true'. Similarly, 'information' is to be explained in terms of the verb 'inform'. Information, we might say, is what is provided when somebody is informed of something. If this is to be a useful pronouncement, we should be able to explain what it is to inform somebody without appeal to phrases like 'to convey information', but this is easily done. To inform someone is to bring them to know something (that they did not already know).

Now, I shall not be seeking to present a comprehensive overview of the different uses of the terms 'information' or 'inform', nor to exhibit the feel for philosophically charged nuance of an Austin. It will suffice for our purposes merely to focus on some of the broadest features of the concept, or rather, concepts, of information.

The first and most important of these features to note is the distinction between the everyday concept of information and technical notions of information, such as that deriving from the work of Shannon (1948). The everyday concept of information is closely associated with the concepts of knowledge, language and meaning; and it seems, furthermore, to be reliant in its central application on the the prior concept of a person (or, more broadly, language user) who might, for example, read and understand the information; who might use it; who might encode or decode it.

By contrast, a technical notion of information is specified using a purely mathematical and physical vocabulary and, *prima facie*, will have at most limited and deriviative links to semantic and epistemic concepts[3].

A technical notion of information might be concerned with describing correlations and the statistical features of signals, as in communication theory with the Shan-

[2] An illuminating discussion of mass, count and abstract nouns may be found in Rundle (1979, §§27-29).

[3] For discussion of Dretske's opposing view, however, see below, Section 1.5.

non concept, or it might be concerned with statistical inference (e.g. Fisher, 1925; Kullback and Leibler, 1951; Savage, 1954; Kullback, 1959). Again, a technical notion of information might be introduced to capture certain abstract notions of structure, such as complexity (algorithmic information, Chaitin (1966); Kolmogorov (1965); Solomonoff (1964)) or functional rôle (as in biological information perhaps, cf. Jablonka (2002) for example[4]).

In this thesis our concern is information theory, quantum and classical, so we shall concentrate on the best known technical concept of information, the Shannon information, along with some closely related concepts from classical and quantum information theory. The technical concepts of these other flavours I mention merely to set to one side[5].

With information in the everyday sense, a characteristic use of the term is in phrases of the form: 'information *about* p', where p might be some object, event, or topic; or in phrases of the form: 'information *that* q'. Such phrases display what is often called *intentionality*. They are directed towards, or are about something (which something may, or may not, be present). The feature of intentionality is notoriously resistant to subsumption into the bare physical order.

As I have said, information in the everyday sense is intimately linked to the concept of knowledge. Concerning information we can distinguish between possessing information, which is to have knowledge; acquiring information, which is to gain knowledge; and containing information, which is sometimes the same as containing knowledge[6]. Acquiring information is coming to possess it; and as well as being acquired by asking, reading or overhearing, for example, we may acquire information via perception. If something is said to contain information then this is because it provides, or may be used to provide, knowledge. As we shall presently see, there are at least two importantly distinct ways

[4]N.B. To my mind, however, Jablonka overstates the analogies between the technical notion she introduces and the everyday concept.

[5]Although it will be no surprise that one will often find the same sorts of ideas and mathematical expressions cropping up in the context of communication theory as in statistical inference, for example. There are also links between algorithmic information and the Shannon information: the average algorithmic entropy of a thermodynamic ensemble has the same value as the Shannon entropy of the ensemble (Bennett, 1982).

[6]Containing information and containing knowledge are not always the same: we might, for example say that a train timetable contains information, but not knowledge.

in which this may be so.

It is primarily a person of whom it can be said that they possess information, whilst it is objects like books, filing cabinets and computers that contain information (cf. Hacker, 1987). In the sense in which my books contain information and knowledge, I do not. To contain information in this sense is to be used to store information, expressed in the form of propositions[7], or in the case of computers, encoded in such a way that the facts, figures and so on may be decoded and read as desired.

On a plausible account of the nature of knowledge originating with Wittgenstein (e.g. Wittgenstein, 1953, §150) and Ryle (1949), and developed, for example by White (1982), Kenny (1989) and Hyman (1999), to have knowledge is to possesses a certain capacity or ability, rather than to be in some state. On this view, the difference between possessing information and containing information can be further elaborated in terms of a category distinction: to possess information is to have a certain ability, while for something to contain information is for it to be in a certain state (to possess certain occurrent categorical properties). We shall not, however, pursue this interesting line of analysis further here (see Kenny (1989, p.108) and Timpson (2000, §2.1) for discussion).

In general, the grounds on which we would say that something contains information, and the senses in which it may be said that information is contained, are rather various. One important distinction that must be drawn is between containing information *propositionally* and containing information *inferentially*. If something contains information propositionally, then it does so in virtue of a close tie to the expression of propositions. For example, the propositions may be written down, as in books, or on the papers *in* the filing cabinet. Or the propositions might be otherwise recorded; perhaps encoded, on computers, or on removable disks. The objects said to contain the information in these examples are the books, the filing cabinet, the computers, the disks.

That these objects can be said to contain information *about* things, derives from the fact that the sentences and symbols inscribed or encoded, possess meaning and hence themselves can be about, or directed towards something. Sentences and symbols, in turn, possess meaning in virtue of their rôle within a framework of language and

[7] Or perhaps expressed pictorially, also.

language users.

If an object A contains information about B[8] in the second sense, however, that is, *inferentially*, then A contains information about B because there exist correlations between them that would allow inferences about B from knowledge of A. (A prime example would be the thickness of the rings in a tree trunk providing information about the severity of past winters.) Here it is the possibility of our *use* of A, as part of an inference providing knowledge, that provides the notion of information *about*[9]. And note that the concept of knowledge is functioning prior to the concept of containing information: as I have said, the concept of information is to be explained in terms of the provision of knowledge.

It is with the notion of containing information, perhaps, that the closest links between the everyday notion of information and ideas from communication theory are to be found. The technical concepts introduced by Shannon may be very helpful in describing and quantifying any correlations that exist between A and B. But note that describing and quantifying correlations does not provide us with a concept of why A may contain information (inferentially) about B, in the everyday sense. Information theory can describe the facts about the existence and the type of correlations; but to explain *why* A contains information inferentially about B (if it does), we need to refer to facts at a different level of description, one that involves the concept of knowledge. A further statement is required, to the effect that: 'Because of these correlations, we can learn something about B'. Faced with a bare statement: 'Such and such correlations exist', we do not have an explanation of why there is any link to information. It is because correlations may sometimes be used as part of an inference providing knowledge, that we may begin to talk about containing information.

While I have distinguished possessing information (having knowledge) from containing information, there does exist a very strong temptation to try to explain the former in terms of the latter. However, caution is required here. We have many metaphors that suggest us filing away facts and information in our heads, brains and minds; but these *are* metaphors. If we think the possession of information is to be explained by our

[8] Which might be another object, or perhaps an event, or state of affairs.

[9] Such inferences may become habitual and in that sense, automatic and un-reflected upon.

containing information, then this cannot be 'containing' in the straightforward sense in which books and filing cabinets contain information (propositionally), for our brains and minds do not contain statements written down, nor even encoded. As we have noted, books, computers, and so on contain information about various topics because they are used by humans (language users) to store information. As Hacker remarks:

> ...we do not *use* brains as we use computers. Indeed it makes no more sense to talk of storing information in the brain than it does to talk of having dictionaries or filing cards in the brain as opposed to having them in a bookcase or filing cabinet. (Hacker, 1987, p.493)

We do not stand to our brains as an external agent to an object of which we may make use to record or encode propositions, or on which to inscribe sentences.

A particular danger that one faces if tempted to explain possessing information in terms of containing it, is of falling prey to the *homunculus fallacy* (cf. Kenny, 1971).

The homunculus fallacy is to take predicates whose normal application is to complete human beings (or animals) and apply them to parts of animals, typically to brains, or indeed to any insufficiently human-like object. The fallacy properly so-called is attempting to argue from the fact that a person-predicate applies to a person to the conclusion that it applies to his brain or *vice versa*. This form of argument is non-truth-preserving as it ignores the fact that the term in question must have a different meaning if it is to be applied in these different contexts.

'Homunculus' means 'miniature man', from the Latin (the diminutive of *homo*). This is an appropriate name for the fallacy, for in its most transparent form it is tantamount to saying that there is a little man in our heads who sees, hears, thinks and so on. Because if, for example, we were to try to explain the fact that a person sees by saying that images are produced in his mind, brain or soul (or whatever) then we would not have offered any explanation, but merely postulated a little man who perceives the images. For exactly the same questions arise about what it is for the mind/brain/soul to perceive these images as we were trying to answer for the whole human being. This is a direct consequence of the fact that we are applying a predicate—'sees'—that applies properly only to the whole human being to something which is merely a part of a human

being, and what is lacking is an explanation of what the term means in this application. It becomes very clear that the purported explanation of seeing in terms of images in the head is no explanation at all, when we reflect that it gives rise to an infinite regress. If we see in virtue of a little man perceiving images in our heads, then we need to explain what it is for him to perceive, which can only be in terms of another little man, and so on.

The same would go, *mutatis mutandis*, for an attempt to explain possession of information in terms of containing information propositionally. Somebody is required to read, store, decode and encode the various propositions, and peruse any pictures; and this leads to the regress of an army of little men. Again, the very same difficulty would arise for attempts to describe possessing information as containing information inferentially: now the miniature army is required to draw the inferences that allow knowledge to be gained from the presence of correlations.

This last point indicates that a degree of circumspection is required when dealing with the common tendency to describe the mechanisms of sensory perception in terms of information reaching the brain. In illustration (cf. Hacker, 1987), it has been known since the work of Hubel and Weisel (see for example Hubel and Wiesel (1979)) that there exist systematic correlations between the responses of groups of cells in the visual striate cortex and certain specific goings-on in a subject's visual field. It seems very natural to describe the passage of nerve impulses resulting from retinal stimuli to particular regions of the visual cortex as visual information reaching the brain. This is unobjectionable, so long as it is recognised that this is not a passage of information in the sense in which information has a direct conceptual link to the acquisition of knowledge. In particular, the visual information is not information for the subject about about the things they have seen. The sense in which the brain contains visual information is rather the sense in which a tree contains information about past winters.

Equipped with suitable apparatus, and because he knows about a correlation that exists, the neurophysiologist may make, from the response of certain cells in the visual cortex, an inference about what has happened in the subject's visual field. But the brain is in no position to make such an inference, nor, of course, an inference of any

kind. Containing visual information, then, is containing information inferentially, and trying to explain a person's possession of information about things seen as their brain containing visual information would lead to a homunculus regress: who is to make the inference that provides knowledge?

This is not to deny the central importance and great interest of the scientific results describing the mechanisms of visual perception for our understanding of how a person can gain knowledge of the world surrounding them, but is to guard against an equivocation. The answers provided by brain science are to questions of the form: what are the causal mechanisms which underlie our ability to gain visual knowledge? This is misdescribed as a question of how information flows, if it is thought that the information in question is the information that the subject comes to possess. One might have 'information flow' in mind, though, merely as a picturesque way of describing the processes of electrochemical activity involved in perception, in analogy to the processes involved in the transmission of information by telephone and the like. This use is clearly unproblematic, so long as one is aware of the limits of the analogy. (We don't want the question to be suggested: so who answers the telephone? This would take us back to our homunculi.)

1.2 The Shannon Information and related concepts

The technical concept of information relevant to our discussion, the Shannon information, finds its home in the context of communication theory. We are concerned with a notion of *quantity* of information; and the notion of quantity of information is cashed out in terms of the resources required to transmit messages (which is, note, a very limited sense of quantity). I shall begin by highlighting two main ways in which the Shannon information may be understood, the first of which rests explicitly on Shannon's 1948 noiseless coding theorem.

1.2.1 Interpretation of the Shannon Information

It is instructive to begin by quoting Shannon:

> The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently these messages have *meaning*...These semantic aspects of communication are irrelevant to the engineering problem. (Shannon, 1948, p.31)

The communication system consists of an information source, a transmitter or encoder, a (possibly noisy) channel, and a receiver (decoder). It must be able to deal with *any* possible message produced (a string of symbols selected in the source, or some varying waveform), hence it is quite irrelevant whether what is actually transmitted has any meaning or not, or whether what is selected at the source might convey anything to anybody at the receiving end. It might be added that Shannon arguably understates his case: in the *majority* of applications of communication theory, perhaps, the messages in question will not have meaning. For example, in the simple case of a telephone line, what is transmitted is not *what is said* into the telephone, but an analogue signal which records the *sound waves* made by the speaker, this analogue signal then being transmitted digitally following an encoding.

It is crucial to realise that 'information' in Shannon's theory is not associated with individual messages, *but rather characterises the source of the messages*. The point of characterising the source is to discover what capacity is required in a communications channel to transmit all the messages the source produces; and it is for this that the concept of the Shannon information is introduced. The idea is that the statistical nature of a source can be used to reduce the capacity of channel required to transmit the messages it produces (we shall restrict ourselves to the case of discrete messages for simplicity).

Consider an ensemble X of letters $\{x_1, x_2, \ldots, x_n\}$ occurring with probabilities $p(x_i)$. This ensemble is our source[10], from which messages of N letters are drawn. We are concerned with messages of very large N. For such messages, we know that typical sequences of letters will contain $Np(x_i)$ of letter x_i, $Np(x_j)$ of x_j and so on. The number of distinct typical sequences of letters is then given by

$$\frac{N!}{Np(x_1)!Np(x_2)!\ldots Np(x_n)!}$$

[10] More properly, this ensemble *models* the source.

and using Stirling's approximation, this becomes $2^{NH(X)}$, where

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i), \tag{1.1}$$

is the Shannon information (logarithms are to base 2 to fix the units of information as binary bits).

Now as $N \to \infty$, the probability of an atypical sequence appearing becomes negligible and we are left with only $2^{NH(X)}$ equiprobable typical sequences which need ever be considered as possible messages. We can thus replace each typical sequence with a binary code number of $NH(X)$ bits and send that to the receiver rather than the original message of N letters ($N \log n$ bits).

The message has been compressed from N letters to $NH(X)$ bits ($\leq N \log n$ bits). Shannon's noiseless coding theorem, of which this is a rough sketch, states that this represents the optimal compression (Shannon 1948). The Shannon information is, then, appropriately called a measure of information because it represents the maximum amount that messages consisting of letters drawn from an ensemble X can be compressed.

One may also make the derivative statement that the information *per letter* in a message is $H(X)$ bits, which is equal to the information of the source. But 'derivative' is an important qualification: we can only consider a letter x_i drawn from an ensemble X to have associated with it the information $H(X)$ if we consider it to be a member of a typical sequence of N letters, where N is large, drawn from the source.

Note also that we must strenuously resist any temptation to conclude that because the Shannon information tells us the maximum amount a message drawn from an ensemble can be compressed, that it therefore tells us the irreducible meaning content of the message, specified in bits, which somehow possess their own intrinsic meaning. This idea rests on a failure to distinguish between a code, which has no concern with meaning, and a language, which does (cf. Harris (1987)).

Information and Uncertainty

Another way of thinking about the Shannon information is as a measure of the amount of information that we *expect* to gain on performing a probabilistic experiment. The Shannon measure is a measure of the uncertainty of a probability distribution as well as serving as a measure of information. A measure of uncertainty is a quantitative measure of the lack of concentration of a probability distribution; this is called an uncertainty because it measures our uncertainty about what the outcome of an experiment completely described by the probability distribution in question will be. Uffink (1990) provides an axiomatic characterisation of measures of uncertainty, deriving a general class of measures, $U_r(\vec{p})$, of which the Shannon information is one (see also Maassen and Uffink 1989). The key property possessed by these measures is Schur concavity (for details of the property of Schur concavity, see Uffink (1990), Nielsen (2001) and Section 2.3.1 below).

Imagine a random probabilistic experiment described by a probability distribution $\vec{p} = \{p(x_1), \ldots, p(x_n)\}$. The intuitive link between uncertainty and information is that the greater the uncertainty of this distribution, the more we stand to gain from learning the outcome of the experiment. In the case of the Shannon information, this notion of how much we gain can be made more precise.

Some care is required when we ask 'how much do we know about the outcome?' for a probabilistic experiment. In a certain sense, the shape of the probability distribution might provide no information about what an individual outcome will actually be, as any of the outcomes assigned non-zero probability can occur. However, we can use the probability distribution to put a *value* on any given outcome. If it is a likely one, then it will be no surprise if it occurs, so of little value; if an unlikely one, it is a surprise, hence of higher value. A nice measure for the value of the occurrence of outcome x_i is $-\log p(x_i)$, a decreasing function of the probability of the outcome. We may call this the 'surprise' information associated with outcome x_i; it measures the value of having observed this outcome of the experiment (as opposed to: not bothering to observe it at all) given that we know the probability distribution for the outcomes[11].

[11] Of course, this is a highly restricted sense of 'value'. It does not, for example, refer to how much

If the information (in this restricted sense) that we would gain if outcome x_i were to occur is $-\log p(x_i)$, then before the experiment, the amount of information we expect to gain is given by the expectation value of the 'surprise' information, $\sum_i p(x_i)(-\log p(x_i))$; and this, of course, is just the Shannon information H of the probability distribution $\vec{p}$. Hence the Shannon information tells us our expected information gain.

More generally, though, any of the measures of uncertainty $U_r(\vec{p})$ may be understood as measures of information gain; and a similar story can be told for measures of 'how much we know' given a probability distribution. These will be the inverses of an uncertainty: we want a measure of the concentration of a probability distribution; the more concentrated, the more we know about what the outcome will be; which just means, the better we can predict the outcome. (To say in this way that we have certain amount of information (knowledge) about what the outcome of an experiment will be, therefore, is not to claim that we have partial knowledge of some predetermined fact about the outcome of an experiment.)

The minimum number of questions needed to specify a sequence

The final common interpretation of the Shannon information is as the minimum average number of binary questions needed to specify a sequence drawn from an ensemble (Uffink 1990; Ash 1965), although this appears not to provide an interpretation of the Shannon information actually independent of the previous two.

Imagine that a long sequence N of letters is drawn from the ensemble X, or that N independent experiments whose possible outcomes have probabilities $p(x_i)$ are performed, but the list of outcomes is kept from us. Our task is to determine what the sequence is by asking questions to which the guardian of the sequence can only answer 'yes' or 'no'; and we choose to do so in such a manner as to minimize the average number of questions needed. We need to be concerned with the *average* number to rule out lucky guesses identifying the sequence.

might be implied by this particular outcome having occurred, nor to the value of what might be learnt from it, nor the value of what it conveys (if anything); these ideas all lie on the 'everyday concept of information' side that is not being addressed here. The distinction between the surprise information and the everyday concept becomes very clear when one reflects that what one learns from a particular outcome may well be, in fact generally will be, quite independent of the probability assigned to it.

If we are trying to minimize the average number of questions, it is evident that the best questioning strategy will be one that attempts to rule out half the possibilities with each question, for then whatever the answer turns out to be, we still get the maximum value from each question. Given the probability distribution, we may attempt to implement this strategy by dividing the possible outcomes of each individual experiment into classes of equal probability, and then asking whether or not the outcome lies in one of these classes. We then try and repeat this process, dividing the remaining set of possible outcomes into two sets of equal probabilities, and so on. It is in general not possible to proceed in this manner, dividing a finite set of possible outcomes into two sets of equal probabilities, and it can be shown that in consequence the average number of questions required if we ask about each individual experiment in isolation is greater than or equal to $H(X)$. However, if we consider the N repeated experiments, where N tends to infinity, and consider asking joint questions about what the outcomes of the independent experiments were, we can always divide the classes of possibilities of (joint) outcomes in the required way. Now we already know that for large N, there are $2^{NH(X)}$ typical sequences, so given that we can strike out half the possible sequences with each question, the minimum average number of questions needed to identify the sequence is $NH(X)$. (These last results are again essentially the noiseless coding theorem.)

It is not immediately obvious, however, why the minimum average number of questions needed to specify a sequence should be related to a notion of information. (Again, the tendency to think of bits and binary questions as irreducible meaning elements is to be resisted.) It seems, in fact that this is either just another way of talking about the maximum amount that messages drawn from a given ensemble can be compressed, in which case we are back to the interpretation of the Shannon information in terms of the noiseless coding theorem, or it is providing a particular way of characterising how much we stand to gain from learning a typical sequence, and we return to an interpretation in terms of our expected information gain.

1.2.2 More on communication channels

So far we have concentrated on only one aspect of describing a communication system, namely, on characterising the information source. The other important task is to characterise the communication channel.

A channel is defined as a device with a set $\{x_i\}$ of input states, which are mapped to a set $\{y_j\}$ of output states. If a channel is noisy then this mapping will not be one-to-one. A given input could give rise to a variety of output states, as a result of noise. The basic type of channel—the *discrete memoryless channel*—is characterised in terms of the conditional probabilities $p(y_j|x_i)$: given that input x_i is prepared, what is the probability that output y_j will be produced?

If the distribution, $p(x_i)$, for the probability with which the various inputs will be prepared is also specified, then we may calculate the joint distribution $p(x_i \wedge y_j)$. We may consider which input state is prepared on a given use of the channel to be a random variable X, with $p(X = x_i) = p(x_i)$; which output produced to be a random variable Y, $p(Y = y_j) = p(y_j)$; and we may consider also the joint random variable $X \wedge Y$, where $p(X \wedge Y = x_i \wedge y_j) = p(x_i \wedge y_j)$.

The joint distribution $p(x_i \wedge y_j)$ allows us to define the joint uncertainty

$$H(X \wedge Y) = -\sum_{i,j} p(x_i \wedge y_j) \log p(x_i \wedge y_j), \tag{1.2}$$

and an important quantity known as the 'conditional entropy':

$$H(X|Y) = \sum_j p(y_j)\big(-\sum_i p(x_i|y_j) \log p(x_i|y_j)\big). \tag{1.3}$$

The scare quotes are significant, as this quantity is not actually an entropy or uncertainty itself, but is rather the *average* of the uncertainties of the conditional distributions for the input, given a particular Y output. It measures the average of how uncertain someone will be about the X value when they have observed an output Y value.

As Uffink (1990, §1.6.6) notes, it pays to attend to the fact that $H(X|Y)$ is not a

measure of uncertainty. It is easy to show (e.g. Ash, 1965, Thm.1.4.3–5) that

$$H(X|Y) \leq H(X), \text{ with equality } \textit{iff } X \text{ and } Y \text{ are independent;} \tag{1.4}$$

and it is often held that this is a particularly appealing feature of the Shannon measure of information because it captures the intuitive idea that by learning the value of Y, we gain some information about X, therefore our uncertainty in the value of X should go down (unless the two are independent). Thus, Shannon describes the inequality (1.4) as follows:

> The uncertainty of X is never increased by knowledge of Y. It will be decreased unless Y and X are independent events, in which case it is not changed. (Shannon, 1948, p.53)

But this description is highly misleading. As Uffink remarks, one's uncertainty certainly *can* increase following an observation: increasing knowledge need not lead to a decrease in uncertainty. This is well illustrated by Uffink's 'keys' example: my keys are in my pocket with a high probability, if not, they could be in a hundred places all with equal (low) probability. This distribution is highly concentrated so my uncertainty is low. If I look, however, and find that my keys are not in my pocket, then my uncertainty as to their whereabouts increases enormously. An increase in knowledge has led to an increase in uncertainty.

This does not conflict with the inequality (1.4), of course, as the latter involves an average over post-observation uncertainties. Uffink remarks, against Jaynes (1957, p.186) for example, that

> ...there is no paradox in an increase of uncertainty about the outcome of an experiment as a result of information about its distribution. The confusion is caused by a liberal use of the multifaceted term information, and also by the deceptive name of conditional entropy for what is actually an average of the entropies of conditional distributions. (Uffink, 1990, p.83)

To see why the conditional entropy is important, consider a very large number N of repeated uses of our channel. There are $2^{NH(X)}$ typical X (input) sequences that could arise, $2^{NH(Y)}$ typical output sequences that could be produced, and $2^{NH(X\wedge Y)}$ typical

sequences of pairs of X, Y values that could obtain. Suppose someone observes which Y sequence has actually been produced. If the channel is noisy, then there is more than one input X sequence that could have given rise to it. The conditional entropy measures the number of possible input sequences that could have given rise to the observed output (with non-vanishing probability).

If there are $2^{NH(X \wedge Y)}$ typical sequences of pairs of X, Y values, then the number of typical X sequences that could result in the production of a given Y sequence will be given by

$$\frac{2^{NH(X \wedge Y)}}{2^{NH(Y)}} = 2^{N(H(X \wedge Y) - H(Y))}.$$

Due to the logarithmic form of H, $H(X \wedge Y) = H(Y) + H(X|Y)$, and it follows that the number of input sequences consistent with a given output sequence will be $2^{NH(X|Y)}$.

Shannon (1948, §12) points out that this means that if one is trying to use a noisy channel to send a message, then the conditional entropy specifies the number of bits per letter that would need to be sent by an auxiliary *noiseless* channel in order to correct all the errors that have crept into the transmitted sequence, as a result of the noise. If input and output states are perfectly correlated, i.e., there is no noise, then obviously $H(X|Y) = 0$.

Another most important quantity is the *mutual information*, $H(X : Y)$, defined as

$$H(X : Y) = H(X) - H(X|Y). \tag{1.5}$$

It follows from Shannon's *noisy coding theorem* (1948) that the mutual information $H(X : Y)$ governs the rate at which information may be sent over a channel with input distribution $p(x_i)$, with vanishingly small probability of error.

The following sorts of heuristic interpretations of the mutual information may also be given: With a noiseless channel, an output Y sequence would contain as much information as the input X sequence, i.e., $NH(X)$ bits. If there is noise, it will contain less. We know, however, that $H(X|Y)$ measures the number of bits per letter needed to correct an observed Y sequence, therefore the amount of information this sequence actually contains will be $NH(X) - NH(X|Y) = NH(X : Y)$ bits.

Or again, we can say that $NH(X:Y)$ provides a measure of the amount that we are able learn about the identity of an input X sequence from observing the output Y sequence: There are $2^{NH(X|Y)}$ input sequences that will be compatible with an observed output sequence, and the size of this group, as a fraction of the total number of possible input sequences, may be used a measure of how much we have narrowed down the identity of the X sequence by observing the Y sequence. This fractional size is

$$\frac{2^{NH(X|Y)}}{2^{NH(X)}} = \frac{1}{2^{NH(X:Y)}},$$

and the smaller this fraction—hence the greater $H(X:Y)$—the more one learns from learning the Y sequence.

The most important interpretation of the mutual information does derive from the noisy coding theorem, however. Consider, as usual, sequences of length N, where N is large; the input distribution to our channel is $p(x_i)$. Roughly speaking, the noisy coding theorem tells us that it is possible to find $2^{NH(X:Y)}$ X sequences of length N (code words) such that on observation of the Y sequence produced following preparation of one of these code words, it is possible to infer which X sequence was prepared, with a probability of error that tends to zero as N tends to infinity (Shannon, 1948). So if we were now to consider an information source W, producing messages with an information of $H(W) = H(X:Y)$, each output sequence of length N from this source could be associated with an X code word, and hence messages from W be sent over the channel with arbitrarily small error as N is increased[12].

The *capacity*, $\mathcal{C}$, of a channel is defined as the supremum over all input distributions $p(x_i)$ of $H(X:Y)$. The noiseless coding theorem states that given a channel with capacity $\mathcal{C}$ and an information source with an information of $H \leq \mathcal{C}$, there exists a coding system such that the output of the source can be transmitted over the channel with an arbitrarily small frequency of errors.

[12]This result is particularly striking as it is not intuitively obvious that in the presence of noise, arbitrarily good transmission may be achieved without the per letter rate of information transmission also tending to zero. The noisy coding theorem assures us that it can be achieved.

1.2.3 Interlude: Abstract/concrete; technical, everyday

Part of my aim in this chapter has been to deflect the pressure of the question 'What is information?' by following the lead of Austin (and, of course, Wittgenstein[13]) and pointing to the fact that 'information' is an abstract noun: correspondingly we should not seek to illuminate the term by attempting fruitlessly to grasp for something that it corresponds or refers to, but by considering simple examples of its function and in particular considering its relations to grammatically simpler and less mystifying terms like 'inform'.

Now, when turning to information in the technical sense of Shannon's theory, we explicitly do *not* seek to understand this noun by comparison with the verb 'inform'. 'Information' in the technical sense is evidently not derived from a nominalization of this verb. Nonetheless, 'information' remains an abstract, rather than a concrete noun: it doesn't serve to refer to a material thing or substance. In this regard, note that the distinction 'abstract/concrete' as applied to nouns does not map onto a distinction between physical concepts and concepts belonging to other categories. Thus the fact that 'information', in the technical sense of Shannon's theory, may be included as a concept specifiable in physical terms does not entail that it stands for a concrete particular, entity or substance. For example, energy is a paradigmatic physical concept (to use another relevant term, energy is a physical quantity), yet 'energy' is an abstract (mass) noun (akin to a property name). The interesting differences that exist between energy and the technical notion of information as examples of physical quantities deserve further analysis. See Chapter 3, Sections 3.4; 3.6 for some remarks in this direction.

Why my insistence that 'information' in the technical sense remains an abstract noun? Well, consider that two strategies present themselves for providing an answer to the question 'What is information' in the case of information theory. On the first the answer is: what is quantified by the Shannon information and mutual information. On the second it is: what is transmitted by information sources. These different strategies

[13] 'The questions "What is length?", "What is meaning?", "What is the number one?" etc., produce in us a mental cramp. We feel that we can't point to anything in reply to them and yet ought to point to something. (We are up against one of the great sources of philosophical bewilderement: a substantive makes us look for a thing that corresponds to it.)' Wittgenstein (1958, p.1).

provide differing, but complementary answers. Under both, however, 'information' is an abstract noun.

Taking the first strategy, one considers what is quantified by the Shannon information and mutual information. As we have seen, the Shannon information serves to quantify how much messages produced by a source can be compressed and the mutual information quantifies the capacity of a channel (for a particular input source distribution) to transmit messages. But this is evidently not to quantify an amount of stuff (even of some very diaphanous kind); and the amount that messages can be compressed and the capacity of a channel are no more concrete things than the size of my shoe is a concrete thing.

Now consider the second strategy. Recall our earlier quotation from Shannon. There he described the fundamental aim of communication theory as that of reproducing at one point a message that was selected at another point. Thus we might say (very roughly) that in the technical case, information is what it is the aim of a communication protocol to transmit: information (in the technical sense) is what is produced by an information source that is required to be reproduced if the transmission is to be counted a success[14].

However, the pertinent sense of 'what is produced' is not the one pointing us towards the concrete systems that are produced by the source on a given occasion, but rather the one which points us towards the particular *type* (sequence or structure) that these tokens instantiate. But a *type* is not a concrete thing, hence 'information', in this technical sense, remains an abstract noun.

So, for example, if the source X produces a string of letters like the following:

$$x_2x_1x_3x_1x_4 \ldots x_2x_1x_7x_1x_4,$$

say, then the type is the sequence '$x_2x_1x_3x_1x_4 \ldots x_2x_1x_7x_1x_4$'; we might name this 'sequence 17'. The aim is to produce at the receiving end of the communication channel another token of this type. What has been transmitted, though, the information

[14]Note that this formulation is left deliberately open. What counts as successful transmission and therefore, indeed, as what one is trying to transmit, depends upon one's aims and interests in setting up a communication protocol.

transmitted on this run of the protocol, is sequence 17; and this is not a concrete thing.

At this point we may draw an illustrative, albeit partial, analogy with information in the everyday sense. Imagine that I write down a message to a friend on a piece of paper (using declaritive sentences, to keep things simple); one will distinguish in the standard way between the sentence tokens inscribed and what is said by the sentences: the propositions expressed[15]. It is the latter, *what is said*, that is the information (everyday sense) I wish to convey. Similarly with information in the technical sense just described: one should distinguish between the concrete systems that the source outputs and the *type* that this output instantiates. Again, it is the latter that is important; this is the information (technical sense) that one is seeking to transmit.

An important disanalogy between the technical and everyday notions of information now forcibly presents itself: the restatement of a by-now familiar point. In the everyday case, when I have written down my message to my friend, one not only has the sentence tokens and the sentence type they instantiate but also the propositions these sentences express; and again, it is these last that are the information I wish to convey. In the case we have just outlined for the information-theoretic notion of information, though, one only has the tokens produced by the source and the type they instantiate; it is this type that is transmitted, that constitutes the information in the technical sense we have just sketched. The further level, if any, of what various types might mean, or what instances of these types might convey, is not relevant to, or discussed by information theory: the point once more that information in the technical sense is not a semantic notion. Indeed, considered from the point of view of information theory, the output of an information source does not even have any syntactic structure.

1.3 Aspects of Quantum Information

Quantum information is a rich theory that seeks to describe and make use of the distinctive possibilities for information processing and communication that quantum systems provide. What draws the discipline together is the recognition that far from quantum

[15]Note, of course, that the propositions expressed are not to be identified with the sentence types of which the tokens I write are particular instances. (Consider, for example, indexicals.)