University of Pennsylvania
ScholarlyCommons

Technical Reports (CIS) | Department of Computer & Information Science

1-1-2010

Disinformation: A Taxonomy

James M. Alexander
University of Pennsylvania

Jonathan M. Smith
University of Pennsylvania, jms@cis.upenn.edu

Follow this and additional works at: http://repository.upenn.edu/cis_reports

Recommended Citation
Alexander, James M. and Smith, Jonathan M., "Disinformation: A Taxonomy" (2010). *Technical Reports (CIS)*. Paper 920.
http://repository.upenn.edu/cis_reports/920

University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-10-13.

This paper is posted at ScholarlyCommons. http://repository.upenn.edu/cis_reports/920
For more information, please contact repository@pobox.upenn.edu.

Disinformation: A Taxonomy

Abstract
This article outlines steps towards a **disinformation theory**, a simplified and generalized notion of communication that is intended to be, in some way, misleading or deceptive. The model is derived from Shannon's communications model, but with an intentional "noise source" and an *unintended* receiver. Alterations of an image containing a message are used to illustrate a variety of disinformation techniques.

Comments
University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-10-13.

This technical report is available at ScholarlyCommons: http://repository.upenn.edu/cis_reports/920

Disinformation: A Taxonomy*

James M. Alexander and Jonathan M. Smith
CIS Department
University of Pennsylvania

Abstract

This article outlines steps towards a **disinformation theory**, a simplified and generalized notion of communication that is intended to be, in some way, misleading or deceptive. The model is derived from Shannon's communications model, but with an intentional "noise source" and an *unintended* receiver. Alterations of an image containing a message are used to illustrate a variety of disinformation techniques.

1 Introduction: Re-thinking Noisy Communication Channels

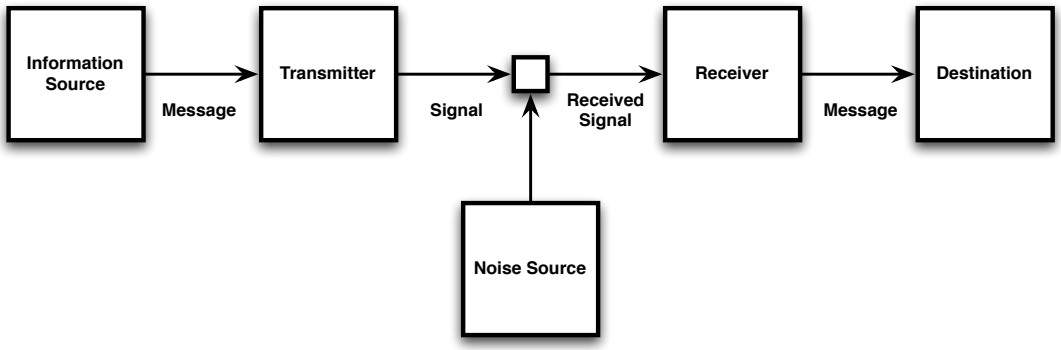

Figure 1: Shannon's Model of a General Communication System

Shannon's simple but powerful model of a communication system[8], illustrated in Figure 1, is sufficiently general to discuss how one might *prevent* communication. A message is intended to be transmitted accurately over a noisy communication channel, and Shannon develops mathematics to characterize the limits of that channel and how to efficiently and effectively communicate over that channel despite its limits.

There is an implicit assumption of *cooperation* between the sender and receiver in this model. In the real world, however, not every communication is a cooperative act: information can be unwittingly transmitted to be eavesdropped by someone we did not intend. The model depicted in Figure 1 can be reinterpreted to capture this case: we simply assume the transmission is not entirely within the control of the owner of the message, and the owner of the receiver has his own motives that may not be very compatible with our own. We consider what countermeasures that the model offers given this situation: if we can't completely prevent the transmission, is there something we can do so that the message is received with some of its content missing? Or might we distort it enough that it is hard to recover or, better yet, it is entirely misleading?

*Supported by NSF CNS-07-16552, University of Pennsylvania Research Foundation and the Olga and Alberico Pompa Professorship

Such a model is applicable to many other problems: cryptographic communications [5], signals intelligence [1], physical camouflage [4], or even poker games [6, 9]. In each case there is information transmission that is not desired: you have an adversary that you don't want figuring out how to interpret the transmission correctly. How can you manipulate the situation to your advantage?

We assume, for this discussion, that although we might be able to manipulate other parts of the abstract communication model, we have absolutely no direct control over the receiver: whatever information makes it through the channel, our adversary is free to interpret as he wishes. That is not to say that there are no useful suppositions that we can make about how the adversary does his work. We can assume, without loss of generality, that the adversary will have a probabilistic model which maps the received signal to all of the possible messages he thinks can receive on the given communication channel. Given a new message, he will measure it according to the parameters of his model, trying to find the most likely match in his model given what he has learned from the message.

We assume for the remainder of the article that we and our adversary both have an equal understanding of how to interpret a given message. That is, the encoding of the message is well-defined and known to both parties. That is, if both parties read the message together, both would agree what it says. Crucially, we *do not* assume that either or both parties believe that the content of the message is accurate. To be more precise, we know which messages we have interfered with and to what extent we have done so, and the adversary may or may not take into account that interference with the communications channel is possible.

Ideally, we would like to mislead the adversary into believing that he has received a message that differs from reality. In order to do a good job of this, we will have to credibly estimate what the adversary's model might look like. If we assume, for illustrative purposes, that we think that the adversary's model has an important two-dimensional subspace, it might be represented graphically something like Figure 2. Each ellipsoid shape delimits the boundaries of a single message, including all possible ways of expressing it. The adversary will map the signal he receives as a point in the model, and interpret it according to the bubble that the message falls inside[1]. Messages are close together in the model if they have similar content. Spaces that are not occupied by messages are allowed by the message encoding, but are excluded from the model for domain-specific semantic reasons: they do not make sense or are excluded by accepted facts. In this example, the correct answer, which we don't want him to know, is inside the shape marked with an asterisk. Our goal is to do our best to ensure that he cannot place the message in the right spot: we must either deprive him of enough information to pick out any one message, or find a means of modifying the message so that he will map it to the *wrong* point.

In order to effectively deceive our adversary with a modified transmission, not just any modification will do. Minimally, the received message must still be valid, meaning it must map to one or more members of the set of possible messages. In addition, the received message must be *plausible*: we must have good reason to believe that if our adversary interprets the falsified message as we intend, he won't disregard it because he too believes it to be false. One part of overcoming this obstacle is to minimize any overt signs of tampering. A much, much harder part of this task is framed well by the following frequently-cited passage from the *The Art of War* by Sun Tzu:

> Know the enemy and know yourself; in a hundred battles you will never be in peril. When you are ignorant of the enemy, but know yourself, your chances of winning or losing are equal. If ignorant both of your enemy and yourself, you are certain in every battle to be in peril.

No matter how sophisticated a theory we develop, there is no substitute for the hard work of gathering all of the knowledge we can about our adversary: we are, after all, trying to guess how he thinks. We find

[1]This particular 2D model was constructed to have a variety of topological densities, but other than that, the details have no particular significance. We are using a concrete example to illustrate the strategies that we might use if this was based upon our best estimate of the adversary's knowledge base.

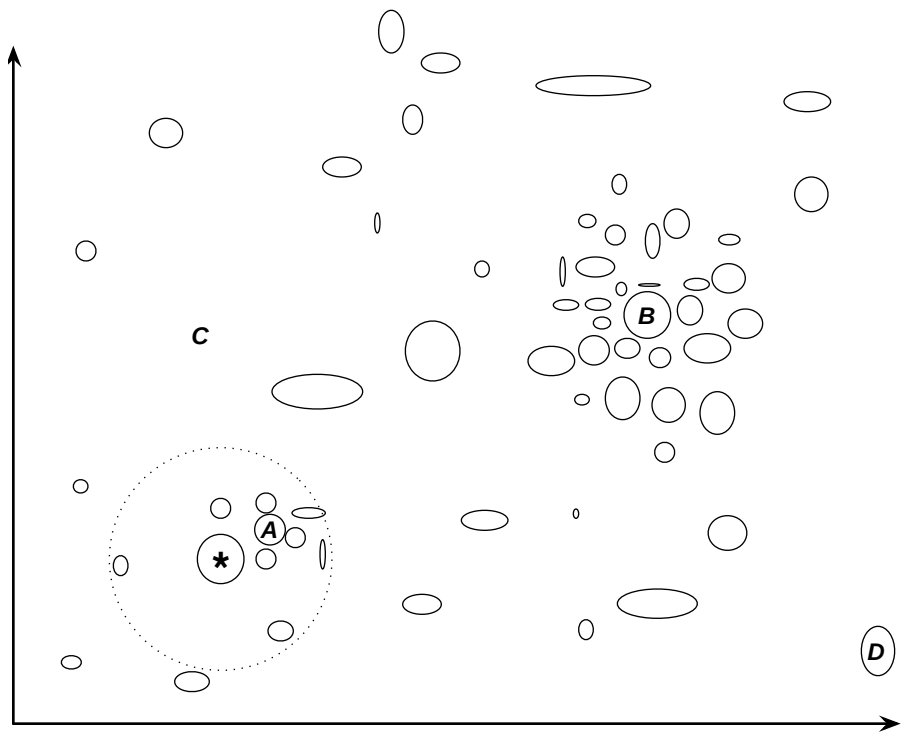

Figure 2: An illustration of an adversary attempting to classify our modified transmission using a 2-dimensional slice of some larger model. Each elliptical region is the subspace occupied by one message. The correct message is the circle marked with an asterisk. The dotted line indicates how aggressive noise-injection might affect his model. The letters indicate various strategies for disinformation, which we explain in detail in the article text.

it useful, therefore, to reason about the model in Figure 2 as if it its topology is influenced by our domain-specific knowledge of the adversary at least as much as the physical constraints of the communication channel itself. We will articulate more about how this might work in the remainder of this article in the context of a running example, introduced in the next section.

2 The Saboteur

The development of disinformation theory took place in the context of the problem of how to prevent identification of faces by automatic face recognition systems. Motivation for why preventing said recognition process might be desirable, as well as how disinformation theory is practically applied to this problem, is addressed in detail in Alexander's dissertation [2]. Broadly, the problem breaks down into two major parts. The first half is how information that uniquely identifies an individual (the message) is encoded in a facial image (the transmission). We put this problem aside for the current paper, and consider the second half: even if we fully understand the information content of the message, do we know how to disrupt it? To get better clarity on this question, we also put aside face recognition in favor of a simpler working problem.

Imagine that there is a war underway, one that is geographically distributed over a large area, but in a pre-radio era. In order to conduct the war, there needs to be communication among far-flung commanders. The state-of-the-art communication technology is written communiqués in plain text on paper, which are carried between the communicating parties by trusted courier. The messages are authenticated with a set of

closely-held paper embossers[2]. You are a undercover spy whose task it is to limit the effectiveness of that communication, preferably without arousing too much suspicion: you want to disrupt as many messages as possible without being caught. For instance, a series of messages that go entirely missing, while completely effective on a per-message basis, is likely to attract unwanted attention rather quickly.

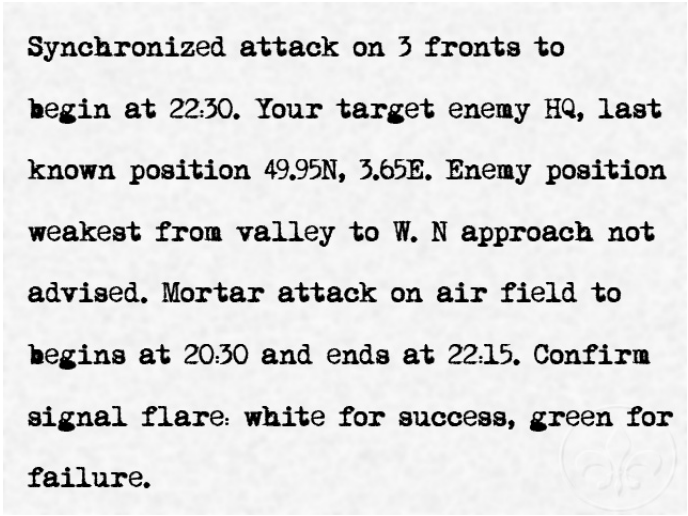
Synchronized attack on 3 fronts to begin at 22:30. Your target enemy HQ, last known position 49.95N, 3.65E. Enemy position weakest from valley to W. N approach not advised. Mortar attack on air field to begins at 20:30 and ends at 22:15. Confirm signal flare: white for success, green for failure.

Figure 3: Our paper-channel message in its original, not-tampered-with form.

In terms of Shannon's model, we define the communications medium to be the typewritten message as carried by the courier. The transmitter includes the chain of people that convey the message verbally and the process of encoding the message onto paper with a typewriter. Once it has left the typewriter, we say that it has been transmitted. Once it had been transmitted, there are opportunities to degrade or modify the content by injecting noise into the communication medium[3].

We discount manipulating the courier in any way due to high risk of detection, and instead focus on the vulnerabilities of the paper component of the medium. What tools do we have available in terms of the model? If we assume, as stated earlier, that we do not have control over the receiving apparatus, that leaves the transmitter and the noise source. In both cases, what we want to do is take full advantage of our knowledge of the limits of the communication medium, overwhelming the theoretical capacity of the channel so that information actually gets lost or, better yet, looks like it has entirely different content than intended. We consider several possible methods of corrupting this communication channel in turn.

In terms of modeling the plausibility of our manipulations, in addition to minimizing the evidence of our manipulations, we need to make sure that the message is believable while still giving some tangible advantage to our side. After all, there is no sense risking changing a message if that change is of no value to us! Relevant knowledge for making the changes seem plausible include recent events in the battle, similarity to past directives, consistency with past behavior of specific commanders, and compatibility with all known current conditions.

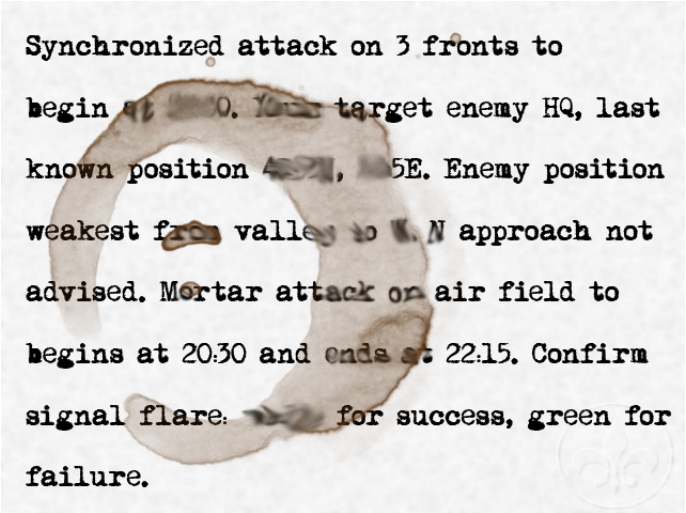

Figure 4: The same message after an unfortunate encounter with a leaky cup of coffee.

3 Destructive Disinformation: Redaction

The simplest, easy-to-accomplish method that we have conceived of is to inject noise in a subset of the signal which we choose to our advantage. We reduce the information content of the message, or, in information theoretic terms, we are increasing its entropy. We do not want to destroy the message outright; instead, we want to damage the subset of the information we believe is the most important. One way of doing this is illustrated in Figure 4. For obvious reasons, we refer to this method as **redaction**. A leaky cup of coffee placed strategically on the message before the courier has it (or indeed after the courier has delivered it, but before it has been read), can remove or severely degrade key parts of the message. Other tools that could cause similar damage to our paper medium include a strategic splash with muddy water or exposure to heat at key positions. In Figure 2, the expected effect of this targeted noise injection is shown as the dotted-line circle - the information content of the message has been reduced, so it could have several possible matches in the receiver's model. In the radio domain [7], this method is akin to intermittent jamming. In a facial photograph, this is like inflicting damage on strategically chosen parts of the image.

The problem with this method, of course, is that the loss of information is completely obvious to the receiver. He knows what pieces of information are missing from the message, doesn't waste a lot of time trying to recover it, and perhaps can reconstruct partial knowledge about the missing pieces based on information that did come though, context, and prior knowledge. The other problem with the obviousness of this technique is that if too many messages are delivered with damage that focuses on only the critical parts of the message, it will be clear that a saboteur is at work. In the next section, we consider a more subtle class of message manipulation, one which is harder to detect and is potentially much more powerful.

[2]This somewhat contrived problem is meant to be more explanatory than realistic. We are well aware that more or less effective cryptographic techniques have been available throughout the history of warfare.

[3]One could make a compelling argument for making a different assignment of the parts of the communication system to the components the abstract channel - we are making a specific choice for clarity of discussion.

4 Constructive Disinformation: Airbrushing

What if, instead of just destroying the key information in the message, we were able to replace parts of it with false, but convincing, information? If we are able to do this job well, not only might the presence of tampering be less obvious, we might be able to convince the receiver that they have received an entirely different message. We are still trying to overflow the channel capacity, but we are doing so in such a way that the affected words *look* like they could be legitimate. We refer to this process as **airbrushing**. Note that forging messages takes us outside the descriptive capability of conventional information theory: we are attempting to fool the receiver into believing that the communication system is behaving normally, when, in fact, it has failed. One cannot characterize this as an increase in entropy as with redaction. It is possible for the *apparent* entropy of the message to change in any way: it can even decrease.

Once we have decided to employ signal airbrushing, there are several strategies we can try: consider Figure 2 once again, focusing on the point marked A. In this variant, we change the shape of the message in a relatively small way, such that we think it will be topologically near the correct message in the receiver's model. That is, we are changing a message that was real, and we can therefore presume its plausibility. We try to preserve its plausibility by leaving most of the message intact, changing it into a nearby message. We illustrate such a modified message in Figure 5. In this case, perhaps the replacement bearings are consistent with historical knowledge about deployments at this position, so the changed message might be even more plausible that the real one was. We are aiming our deception at the highest density cluster of messages that is still a small change: We call this strategy **local crowd-blending**. This change is small, but if chosen well, but might have a big effect on the battle to come.

The imperfections caused by how the change was accomplished might be overlooked, at least once, as a genuine error correction on the part of the typist. Our chances of successful deception are particularly good if we are able to increase the plausibility of the message as desired.

With a facial image, the equivalent of this technique is to use digital retouching methods to reshape or reposition objectively important features. Examples of such methods are given by Alexander [2].

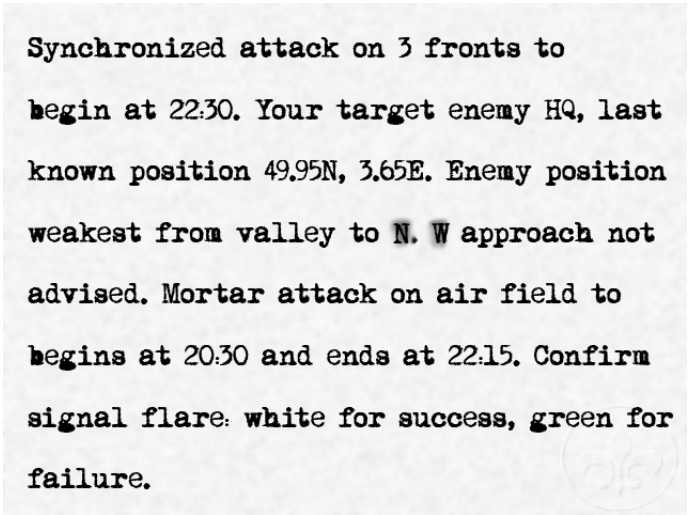
Synchronized attack on 3 fronts to begin at 22:30. Your target enemy HQ, last known position 49.95N, 3.65E. Enemy position weakest from valley to N. W approach not advised. Mortar attack on air field to begins at 20:30 and ends at 22:15. Confirm signal flare: white for success, green for failure.

Figure 5: The message with a small, key piece of information modified by our spy.

If the stakes are higher, we might wish to make a bolder airbrushing attempt: consider position B in Figure 2. The message has been modified extensively, so it is much farther away from the original. An example of such a change might be illustrated by Figure 6. Ideally, if we are going to make such large

modifications, we would like them to lie in one of the most densely populated regions of the receiver's model. We call this strategy **global crowd-blending**. In the case of our running example, perhaps this choice of flare colors and the timing of the mortar attacks are in accord with the most frequent past events, but are critically wrong for the upcoming battle. It could be that the commanders have successfully employed a strategy of moving their forces during bombardment as an element of surprise. If the artillery gunners were aware of the direction of approach of their own forces, they could adjust their aim accordingly. Of course, in this case, that was not the plan, so if this message is plausible, we have led our enemy into a very nasty trap.

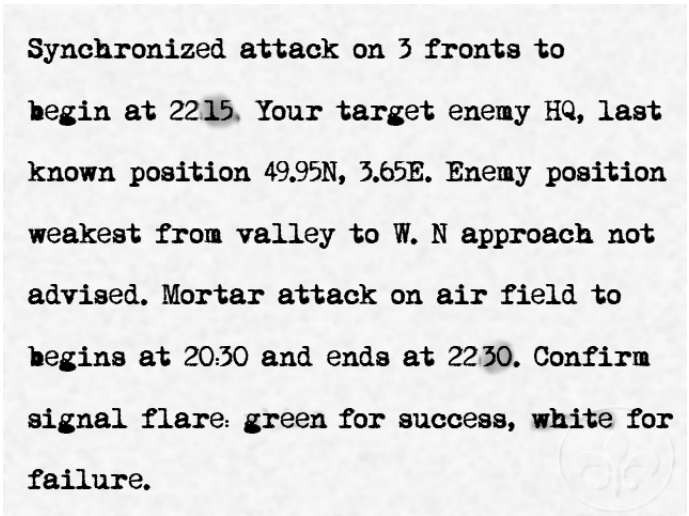
Synchronized attack on 3 fronts to begin at 2215. Your target enemy HQ, last known position 49.95N, 3.65E. Enemy position weakest from valley to W. N approach not advised. Mortar attack on air field to begins at 2030 and ends at 2230. Confirm signal flare: green for success, white for failure.

Figure 6: The message with more substantial modifications.

Of course, no matter how plausible the resulting message, the larger changes cause more obvious artifacts of tampering, making it easier for our adversary or detect our subterfuge. There are ways to improve our chances, however. Consider Figure 7. In this attempt to deceive, we have made the same changes as 6, but are also using redaction as a form of misdirection: we make liberal use of a lighter in the hopes of covering up some the evidence of tampering. Notice we have made sure that the worst damage is confined to parts of the message that are of lesser importance, and the receiver might consider themselves lucky that the key parts are perfectly readable, missing the more subtle tampering that was our real goal.

A variant of this technique, which is difficult to actualize with our paper message channel, is to use two separate levels of interference. One level is meant to be detected: a deliberate, but somewhat clumsy, attempt at disinformation. This level serves as misdirection for more subtle disinformation, which might go undetected, and might be more valuable for our larger strategy.

Instead of making difficult-to-hide modifications, another very valuable airbrushing technique we might try is inserting *additional* distractor information into the message, as illustrated by Figure 8. This method might push us into the vicinity of D in the receiver's model as shown in Figure 2. In this case, we mean that the target message is a statistical outlier, meaning that this is a very unlikely message to be sent, but we still think that it is a message that occurs in their model. Perhaps commanders have heard of retreat being signaled by a flare, but have not actually received such orders themselves. It might strain credulity a bit too far, or might get reluctantly accepted. Certainly, more than one or two unusual messages like these would get accepted without confirmation or investigation. We refer to this as the **curve ball** strategy, and could certainly be accomplished by airbrushing as well.

The last possibility shown in 2, C, is a statistical outlier that is not likely to be anywhere near any

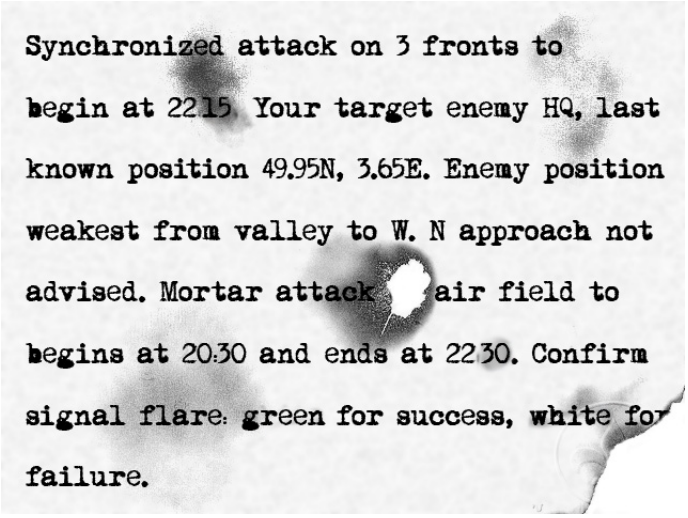
Synchronized attack on 3 fronts to
begin at 22:15. Your target enemy HQ, last
known position 49.95N, 3.65E. Enemy position
weakest from valley to W. N approach not
advised. Mortar attack on air field to
begins at 20:30 and ends at 22:30. Confirm
signal flare: green for success, white for
failure.

Figure 7: The message with the same modifications as Figure 6, but with some noise added to distract from some of the evidence of tampering.

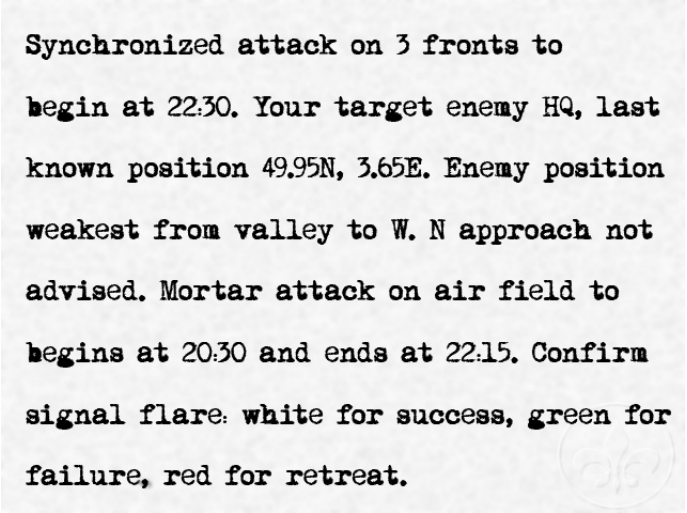
Synchronized attack on 3 fronts to
begin at 22:30. Your target enemy HQ, last
known position 49.95N, 3.65E. Enemy position
weakest from valley to W. N approach not
advised. Mortar attack on air field to
begins at 20:30 and ends at 22:15. Confirm
signal flare: white for success, green for
failure, red for retreat.

Figure 8: A possibly harder-to-detect attempt at inserting disinformation into a message.

match in the model. This might be a message that has been modified to include implausible, outlandish disinformation: the intervention of Martian forces, perhaps. We have not fully worked up this case as a full-fledged example since we don't think it is often a viable strategy (unless you know your adversary to be very gullible). In terms of our face recognition problem, such an implausible outlier might be illustrated by the digital insertion of a third eye.

We conclude this section with one final disinformation method, illustrated by Figure 9. The idea is that we corrupt the message before it ever passes through the typewriter, for instance by passing false information verbally to one or more of the key players. This is, strictly speaking, not the same class of attack: we have taken control of the transmitter rather than manipulating the noise source. If we can pull it off, this method might be very risky indeed, but it has the distinct advantage of producing a transmission that is free of any

evidence of tampering: there have been no changes at the transmission level. The message is totally genuine: it is just not the message that was intended.

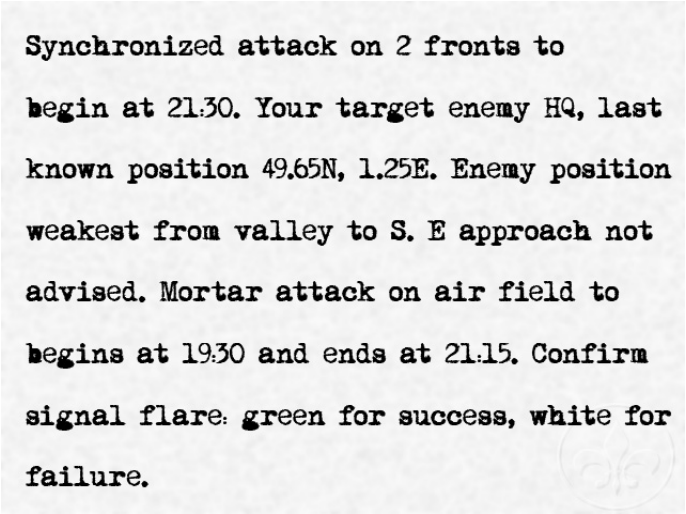
Synchronized attack on 2 fronts to begin at 21:30. Your target enemy HQ, last known position 49.65N, 1.25E. Enemy position weakest from valley to S. E approach not advised. Mortar attack on air field to begins at 19:30 and ends at 21:15. Confirm signal flare: green for success, white for failure.

Figure 9: An ideal disinformation method: insert the false information earlier in the process, and the message will lack detectable artifacts.

5 Future Developments

As shown by Alexander [2], thinking about this simpler communication scenario, even in its relatively abstract form, has been a very helpful tool. It has led us to concrete strategies for disguising faces, and has exposed some of the problems we might encounter in their use in real-world applications.

We additionally believe that there is a lot more depth to be plumbed in this model. We are causing deliberate and precise failures of a communication system: whether the message appears to have been tampered with or not, the intended message is lost to the receiver. We have therefore moved beyond what information theory was intended to describe. In terms of formalizing the model, what is really needed is a framework for reasoning about what our adversary may or may not know, and a method of characterizing how likely our tampering is to go undetected. This moves us more into the realm of behavioral game theory [3], a variant of game theory that tries to capture the psychological element into how human beings make strategic decisions. Such an addition would also be useful for thinking through more complicated scenarios where, for instance, our adversary has partial information about our efforts at sabotaging his communication channel, and might make an effort to communicate steganographically, or perhaps try to goad us into exposing ourselves with messages specifically crafted as bait. We leave these refinements of disinformation theory for future work.

References

[1] Richard J. Aldrich. *Intelligence and the War against Japan: Britain, America and the Politics of Secret Service*. Cambridge University Press, 2000.

[2] James Michael Alexander. *MASKS: Maintaining Anonymity by Sequestering Key Statistics*. PhD thesis, University of Pennsylvania, 2009.

[3] Colin F. Camerer. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press, 2003.

[4] Peter Forbes. *Dazzled and Deceived: Mimicry and Camouflage*. Yale University Press, 2009.

[5] David Kahn. *The Codebreakers*. Macmillan, 1967.

[6] Kevin D. Mitnick. *The Art of Deception: Controlling the Human Element of Security*. Wiley, 2002.

[7] Richard A. Poisel. *Modern Communications Jamming Principles and Techniques*. Artech House, 2003.

[8] Claude E. Shannon and Warren Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, 1949.

[9] David Sklansky. *The Theory of Poker: A Professional Poker Player Teaches You How to Think Like One*. Two Plus Two Publishing, 1994.