To model *learned helplessness:* is depression a maladaptive response of a rational system?

David C. Moffat
Digital Design Technologies
Glasgow Caledonian University, UK.
D.C.Moffat@gcu.ac.uk

Abstract.

Depression is a complex phenomenon, but in a simplified form may be induced in laboratory animals. The response called "learned helplessness" has been associated with depressive symptoms, and can be caused by electrocuting caged dogs. There are hypotheses about the kind of cognitive processes involved, but there is also criticism that this theory does not account for some aspects of depression, including the way that people generalise their feelings beyond the stressful situation that caused their depression.

This paper develops a simple model for the learning processes which might be involved in learned helplessness. It suggests an account for the growth of pessimism, which is characteristic of depression.

1 Background to learned helplessness

Depression is seen as a mental illness, by the sufferer and by observers. Others see it as a non-rational mental state, in which the sufferer misperceives or misinterprets the world. However it might alternatively be seen as a half-rational response to events. One model of depression has been produced in laboratories with animals, and was called "learned helplessness" by Seligman [4]. In this paradigm, dogs are typically electrocuted while prevented from being able to escape. Later, when they are able to escape, they do not do so but merely whimper and tolerate the electrocution until it is over. These cruel experiments were originally done in order to investigate depression. The response of learned helplessness was thought to generalise from dogs in cages to people in stressful life situations [4, 5, 1].

Some people may doubt that depression should be seen as a disorder, especially as it is common in people who manage to continue to live productive and otherwise normal lives. If there is to be found any rational purpose to depressive illness, then perhaps it should not be called an illness at all? There are even depressed people of great achievement, such as Winston Churchill, who was famously dogged by depression even while saving his country, and civilisation as a whole. However Churchill himself did not enjoy his depression, nor saw much benefit in it. On the contrary, it was a hindrance in his efforts. As he and other sufferers would much rather be without it, I take the position that depression is indeed a disorder, in the sense that we should try to find a treatment for it. In this case the supposed rationality of depression comes into focus. While conceding that it might indeed be a rational response in some way, as yet to be determined, and while the remainder of this paper itself gives some attention to this issue, it does not follow that depression is no disorder. The concept of rationality is itself open to multiple interpretations, for one thing. Furthermore, a basically rational response pattern might turn out to be irrational in certain circumstances.

The very concept of depression is contentious in other ways, too. It is a complex phenomenon if it is indeed a single phenomenon at all. It can have a wide variety of symptoms that have no obvious connection to each other, such as appetite loss, sleeplessness, and anhedonia, or taking no pleasure in life any longer. What could cause such different things: surely nothing simple? Combined with the previous more philosophical issues, the scope or even possibility for computational modeling is not appealing. In order to begin, it is wise to simplify first of all. A relatively simple manifestaion of depression would be one that could be produced reliably in a laboratory, and with simpler animals. This paper thus aims to model *learned helplessness.*

The response of learned helplessness shares many symptoms with depressive illness, or at least it manifests analogs. It could be seen as one type of depression, and for some people might even be seen as a common factor in all depression. Such questions are beyond the scope of this paper to settle, or address in any detail, so here I shall take the position that learned helplessness is a factor that is strong enough to be taken as a core behavioural characteristic of at least some forms of depression. A person in a depressed mood may show a lack of interest in most activities, taking no pleasure in them, lose appetite, be agitated or less active in movements, lack energy, and feel worthless or guilty. Similarly, animals who have learned helplessness are less responsive generally, eat less and lose weight, remain passive in the event of electric shock, and act as if they believe they cannot control their environment.

The way that dogs react when they are no longer restrained, is the most dramatic evidence of the phenomenon of learned helplessness. Although they can now escape the electrocution, they typically do not attempt to do so, and instead remain passively in place, waiting for the electrocution to stop.

This behaviour seems to be so dysfunctional that the response, rather like depression in people, is often thought to be irrational. I use the stronger word "irrational" here to include cases where the sufferer (like Winston Churchill) cannot understand his own depression. In a sense he can see his own behaviour as confused and contradictory, and thus as irrational in some part. To be irrational then, an organism must be capable of rationality, must aim to achieve it,

and fail. Alternatively, behaviour may be merely non-rational if it does not issue from some deliberative or goal-oriented mental process. However, as it is a learned response, there is the possibility that learned helplessness should be said to inherit at least some rationality from the capacity to learn, which is generally adaptive for dogs and of course for humans. In that case, it may be possible to construct a formal AI model of a learning agent, which in some cases can develop something like learned helplessness. To the extent that the system can learn and successfully adapt, it is rational; but if and when it fails, or changes in ways that are maladaptive, then it can be called irrational. It is that kind of predictable failure, when a rational system can be induced to malfunction in some cases, that we aim for in this paper.

2 An elementary learner in the cage

An elementary learning agent that is partially able to model something like the learned helplessness behaviours in the laboratory dogs goes as follows. This is a hypothetical model, in the sense that it has not been implemented. However it is simple enough to show how the behaviour patterns would manifest, without requiring implementation. As such it occupies a middle ground between AI and philosophy, or theoretical speculation. Many AI models have to be implemented and executed in order to understand what they do and how regular or reliable they are. In this case, the model is so simple that implementation is not necessary. While some people might see that as a weakness, or not proper AI, in my view it is a strength, because complicated models are often unconvincing, even if they make good predictions. The model is a basic rule-based system, which is standard fare in AI. Some people might doubt that such a simple architecture can model complex mental processes, but in practice many expert systems of impressive complexity and performance have been built for decades in AI, so the architecture is not as limited as it may at first appear. In any event, it turns out to be equal to the task of this paper.

Firstly we give it a set of rules for the situation, and priorities for them that depend on their confidence or success rate. These priorities can then be modified by experience, to give a basic form of adaptation.

Table 1. Small set of rules for the dog, when electrocuted in the cage

priority	rule
0.8	bark
0.7	attack the cage
0.6	flee the cage
0.5	hide
0.3	submit

The priorities are represented as numbers between zero and one in Table 1, where the larger numbers mean higher priorities.

The basic learning scheme then modifies these rule priorities by reinforcement, so that rules which fail then lose priority. Failure in this context means failing to stop the pain. After some attempts to react in this situation by trying the higher priority rules, the dog will fail, and those rules will lose priority so that they are less likely to be tried again. The first rule to drop will be `bark`, and when its priority falls below 0.7 it will be the `attack` rule that is then top priority. To attack the cage would mean biting the door and attempting to hurt the cage as an enemy and stop its attack on the dog. That rule will also fail, of course, as in the experiment situation there is nothing effective that the dog can do. The next rule will be to flee the cage, and this is when could try to jump over the low barrier, if it is able to in the later part of the experiment when it is no longer restrained.

Although this is a very simple scheme, the rules do not have to operate singly as described. Some of their behaviours are compatible, and so may be done at the same time. For example, the dog might bark while counterattacking. It would not be possible to both attack and hide however, or attack and flee, so those rules are mutually exclusive and only the higher priority rule at the time would fire.

After some time and many attempts the priorities of the top rules fall so low that the initially bottom rule is the top one. This is a rule for submissive behaviour, which in dogs is used to give up the fight, signalling defeat. Within it there are different behaviours, including whimpering or crying, adopting a low posture, crouching down, or turning over onto the back to show the belly and become even more vulnerable, and in extreme cases urinating or defecating. The dogs in the experiment behaved in some of those ways, but it is an extra assumption we make here for the model that their behaviour could be considered *submissive*, and not merely passive. Those dogs that would lie down and whimper or urinate were clearly *doing something:* they were not merely being passive.

One further assumption we make for the model is that some rules do not lose priority, and that the rule for submission is one of those. The rationale for this is that the purpose of submissive behaviour is one of last resort, so that animals will submit when there is nothing else they can do. Clearly, to roll over and expose the belly or throat to an attacker is a desperate act, and should generally only be the last resort. Accordingly, this rule does not weaken when if fails, but keeps its priority and thus stays on top of all the other rules. At this point the rules and their priorities might be as shown in Table 2

Table 2. The rules when the dog has been totally inneffective

priority	rule
0.3	submit
0.25	flee the cage
0.23	bark
0.21	attack the cage
0.18	hide

The dog, as predicted by these rules, would now go into its submissive behaviour and stay at the bottom of the cage, whimpering and urinating, but enduring the electrocution.

3 Is this learned helplessness?

The behaviour shown by the agent (simulated dog) in the cage thus appears to be similar to the way the real dogs would behave when electrocuted. At least, it is so far: but what happens when the barrier is lowered so that the dog could jump over it?

The real dogs continue to manifest their passive behaviour, and do not jump over the barrier to escape the electrocution. The simulated dog would continue to execute its top priority rule when electrocuted again; and that is to submit again. The dog has other rules available to it, one of which would work well: namely, to jump over the barrier (an optional part of the `flee` rule). Although it could escape the pain however, the dog as in Table 2 would not do so because the top priority rule is still the one to `submit`.

This is the root of the strange behaviour observed in real dogs, if the model is correct as far as it goes.

Real dogs do not always continue the pattern of learned helplessness for long, however. Most of them do; but some of them break out of the pattern after some time, when they try to jump over the barrier, and succeed. After that, they are more likely to jump again the next time they are electrocuted. This is as we would expect, but it is only some of the dogs that manage to fully break the pattern; and strangely some of them may occasionally fall back into it again later on. That makes the whole phenomenon of learned helplessness all the more astonishing, of course. When the animal knows that it can escape electrocution after all, and has already done so before, why should it sometimes choose not to?

Some people might not find this behaviour surprising, especially if they are familiar with depressive illness or other emotional disorders, in a professional capacity. But to the untutored eye it can look highly irrational for the dog to want so intensely to escape the pain; to know how to escape it having already done so — and yet to fail to escape. To have a top-priority goal, and have a simple plan to achieve it, and even to have executed the plan already on the previous occasion, but then fail to execute that plan, must surely count as irrational. Depressive illness more generally might not be anything like as irrational as this; but the point of this paper is to examine and try to model such extreme cases, where clearly irrational behaviour results from a generally rational system.

The simulation may be able to account for some of this strange behaviour as well. Firstly, note that the lower priority rules in Table 2 are only a little below the priority of the top rule (for `submit`). Therefore, there is a chance that the other rules might fire occasionally. We only take priority to *bias* the chance of a rule's firing, in proportion to how far it exceeds the priorities of those rules just inferior to it. In this case, there is a chance that the rule to `flee` might fire, and if it does then it will succeed, and its priority will rise again. That will make it more likely to be chosen the next time, and so break the pattern of "learned helplessness" (or of the preference to `submit`).

Secondly, by the same mechanism, it is also possible that the pattern of learned helplessness could re-establish itself, because even if the `flee` rule rises above the `submit` rule, it will not be by much initially, and so the dog might still choose to submit again, occasionally.

4 Conclusion

In order to study depression, psychologists have sought to induce it in laboratory animals, and have found that the response of "learned helplessness" offers a useful experimental analog. Psychologists have suggested that the mechanisms at the base of the response may be a recognition of contingency, and the lack of control the animal perceives that it has [3]. However, there is as far as I am aware no attempt made to model the response in software.

If the model as outlined here is correct, and predicts the puzzling behaviour patterns observed in laboratory animals, then it could form the basis of a model of depression, in people as well as dogs. More work will be needed to achieve that, and more questions to be answered on the way, such as how to account for a fuller set of symptoms of depression, and how to account for the generalisation that appears in depressive illness. Even in the laboratory dogs, it is not clear from the psychological theory regarding contingencies, nor from the learning model presented here, why it should be that dogs and depressed people should lose their appetite, and lose weight.

Relevant recent work in psychology has attempted to draw out the statistical correlations between various depressive symptoms, as reported in the clinical psychology literature [2]. There is also a computational model of that work [6], implemented in the netLogo system [7]. That model and implementation are essentially statistical, embodying the relationships between measured variables (symptoms), and as such can be useful in further theoretical development. The aim of this paper however, is to build a different kind of model, which is *causal* in nature, and proposes internal psychological mechanisms that can explain the observed phenomena (symptoms). As such, further work of the sort shown here might one day be able to predict the statistical relationships reported by Borsboom and Cramer [2].

Evidently, depression is a complex phenomenon, and poorly understood as yet. It may be that computational modeling can make incremental progress in penetrating its mysteries, providing causal models that base the disorder in more generic psychological mechanisms, including learning. The work in this paper is an initial attempt to do so, and to throw some light on the question of the rationality, or not, of depression.

5 Acknowledgements

I thank the thoughtful reviewers of this paper, who made useful suggestions, and gave an interesting range of reaction that was in itself enlightening.

REFERENCES

[1] Marc D. Binder, Nobutaka Hirokawa, and Uwe Windhorst, eds., *Encyclopedia of Neuroscience*, chapter Learned Helplessness, 2123–2123, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[2] Denny Borsboom and Angélique OJ Cramer, 'Network analysis: an integrative approach to the structure of psychopathology', *Annual review of clinical psychology*, **9**, 91–121, (2013).

[3] Steven F. Maier and Martin E. Seligman, 'Learned helplessness: Theory and evidence.', *Journal of Experimental Psychology: General*, **105**(1), 3–46, (1976).

[4] M E P Seligman, 'Learned helplessness', *Annual Review of Medicine*, **23**(1), 407–412, (1972).

[5] Martin E. P. Seligman, *Helplessness: On Depression, Development and Death*, Freeman, 1992.

[6] C.D. Van Borkulo, H.L.J. Van der Maas, D. Borsboom, and A.O.J. Cramer. Netlogo Vulnerability_to_Depression, December 2013. http://ccl.northwestern.edu/netlogo/models/community/Vulnerability_to_Depression. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.

[7] Uri Wilensky. Netlogo. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL., 1999. http://ccl.northwestern.edu/netlogo/.