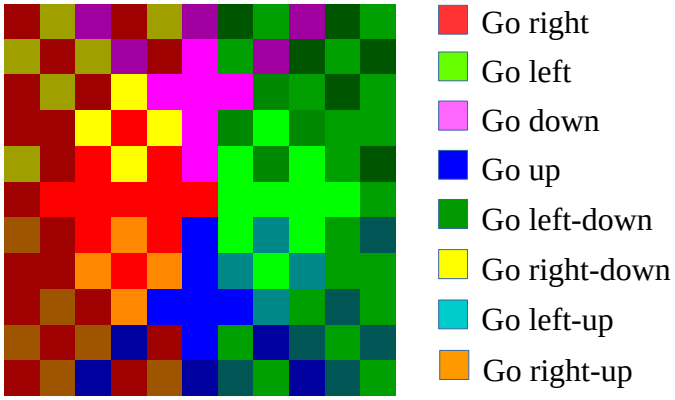
Local Optimal Safety Theory

by Sven Nilsen, 2018

In this paper I represent a decision theory for local optimal safety called "Local Optimal Safety Theory" (LOST) assuming perfect information about the environment. The decision theory is described formally by a higher order function returning the decision algorithm for the optimal safe agent. A LOST agent uses bounded search for the minimum amount of steps required to reach a safe or an unsafe state. The decision at each time step is simply to: Maximize the unsafe minimum distance at next time step when safe, and minimize the safe minimum distance at next time step when unsafe.



A color map displaying decisions made by a pixel world LOST agent programmed to stay inside a circle.

Local Optimal Safety Theory (LOST) is defined as following:

f : S → bool	A function returning `true` when a state is unsafe
c : S → [S]	A decision function returning available states from a given state

lost : (S → bool) × (S → [S]) × nat → (S → S)
steps : (S → bool) × (S → [S]) × nat → nat

`lost` uses `where` ala "secrets" in Dyon for arg-min (Modified Dyon code for clarity in definitions):

```
lost(f: S -> bool, c: S -> [S], n: nat) = \(x: S) = {
    cx := c(x)
    m := if f(x) {min i {steps(\(x) = !f(x), c, n, cx[i])}} else {max i {steps(f, c, n, cx[i])}}
    if is_nan(m) {x} else {cx[where(m)[0]]}
}

steps(f: S -> bool, c: S -> [S], n: nat, x: S) = {
    if f(x) {return 0} else if n <= 0 {return nan()}
    cx := c(x)
    min i {steps(f, c, n - 1, cx[i])} + 1
}
```

LOST agents will seek local optimal safety. A LOST agent is scared of crossing "bridges", because the risk of "falling down" increases temporarily when crossing, in case the agent loses its own mind.

A LOST agent might "panic" when there are few choices (action gradient collapses into local cycles). This happens because there is not enough information to prioritize a choice among equal alternatives.