Carnegie Mellon University

**Center for the Neural Basis
of Cognition (CNBC)**
115 Mellon Institute
4400 Fifth Avenue
Pittsburgh, Pennsylvania 15213-2683

Phone: 412-268-4615
Fax: 412-268-5060
www.cognitiveaxon.com
timothyv@cmu.edu

To the Editors:

We submit our manuscript, *A way around the exploration-exploitation dilemma*, for consideration at *PNAS*.

Since Schultz et al's (*J Neurophys,* 1996; > 4000 citations) seminal report that dopamine neurons behave quantitatively like a reinforcement learning signal, reinforcement learning has been central to the study of learning as well as important to our conception of numerous neurological and psychiatric disorders, including autism and schizophrenia. Reinforcement learning is also enjoying a renaissance in artificial intelligence research, achieving superhuman performance on Atari video games (Minh et al, *Nature* 2015; >6000 citations), chess (Silver et al, Science 2018; >400 citations), and Go (Silver et al, *Nature* 2016; >5000 citations). However, despite this progress all reinforcement learning shares a common dilemma.

The exploration-exploitation dilemma is summarized by a simple question: “Should I exploit an available reward, or explore to try out a new uncertain action?” This dilemma is a problem all decision makers face. It is a problem without a solution. The lack of a solution fundamentally limits our ability to predict optimal naturalistic behavior during foraging and exploration, and to optimally drive learning in artificial agents.

To overcome this field-wide limitation we took a fresh look at the dilemma. Our goal was simple: when one mathematical problem can’t be solved, it’s often good to find another related problem that can be. In this work we restate the dilemma in a way that serves the same aim, but now has a clear and practical solution.

We show that (nearly) any dilemma problem can also be viewed as competition between exploiting known rewards, or exploring to learn a *world model*--a simplified concept of memory borrowed from computer science. We prove this competition can be perfectly solved using the simplest of all value learning algorithms: a deterministic greedy policy. To ensure this solution is as broad as possible we also offer a new minimal and perhaps universal theory of information value that complements--but is independent of--Shannon’s Information Theory.

To a popular audience, we have to emphasize the algorithm we discovered translates to, *“What do you value more right now, getting a reward or learning something new? Pick the biggest”.* This means our analysis has broad but simple application both to scientific research--in fields as diverse as psychology, neuroscience, biology, data science, artificial intelligence, game theory, economics, demography, and political science--and perhaps even directly to everyday life.

With our submission we include the text of the primary manuscript with 3 figures as well as a full supplementary discussion and mathematical appendix. The main text contains about 3000 words and with figures and figure legends will be approximately 4.5 pages in the printed publication. This work is not under consideration for publication elsewhere. Code and data are available at https://github.com/CoAxLab/infomercial.

Erik J Peterson, Ph.D.
Research Fellow
Carnegie Mellon University, Pittsburgh

Tim Verstynen, Ph.D.
Associate Professor
Carnegie Mellon University, Pittsburg