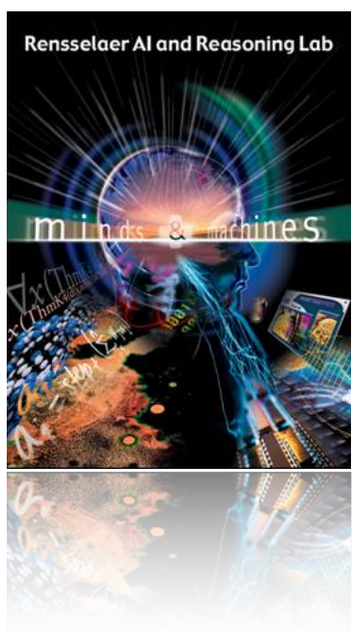
Well, *Zombie* Autonomy is Fearsome — But Can be Tamed

Selmer Bringsjord

Rensselaer AI & Reasoning (RAIR) Lab
Department of Cognitive Science
Department of Computer Science
Lally School of Management & Technology
Rensselaer Polytechnic Institute (RPI)
Troy, New York 12180 USA

IACAP @ U of Delaware
6/22/2015



A “Worrisome” Equation …

that I won’t dignify with any cerebration:

Unpredictability + Explosiveness = Danger, Will Robinson!

Unpredictable(x) + Explosive(x) = Dangerous(x)

Machine Learning should be banned from the relevant domains.

Unfortunately, there are identifications of *unpredictability* with *autonomy* floating around …

“Modern/Selmer DARPA”

$$(\alpha_1^a, \alpha_2^a, \ldots, \alpha_k^a, \alpha_{k+1}^a, \ldots)$$

$$\frac{\mathcal{C}[(\alpha_1^a, \alpha_2^a, \ldots, \alpha_k^a)]}{1}$$

$$\mathcal{H}[(\alpha_1^a, \alpha_2^a, \ldots, \alpha_k^a)]$$

"Modern/Selmer DARPA"

$$(\alpha_1^a, \alpha_2^a, \ldots, \alpha_k^a, \alpha_{k+1}^a, \ldots)$$

$$\left.\frac{\mathcal{C}[(\alpha_1^a, \alpha_2^a, \ldots, \alpha_k^a)]}{\dfrac{1}{\mathcal{H}[(\alpha_1^a, \alpha_2^a, \ldots, \alpha_k^a)]}}\right\}$$

= level of autonomy

(the larger the more autonomous)

= level of unpredictability

A Worrisome Equation

Autonomy + Power = Danger, Will Robinson!

$\forall x : \texttt{Agents}$ Autonomous(x) + Powerful(x) = Dangerous(x)

MMOI

?

Autonomous(x) + Powerful(x) + Means(x) + Opportunity(x) = Dangerous(x)

In other words, we need philosophy and logic.

Desperately…

Shared space in tomorrow's world.

In order to provide a foundation for the new autonomous F 015 Luxury in Motion research vehicle, an interdisciplinary team of experts from Mercedes-Benz has devised a future scenario that incorporates many different aspects of day-to-day mobility. Above and beyond its mobility function, this scenario perceives the motor car as a private retreat that additionally offers an important added value for society at large.

“Anyone who focuses solely on the technology has not yet grasped how autonomous driving will change our society,” emphasises Dr Dieter Zetsche, Chairman of the Board of Management of Daimler AG and Head of Mercedes-Benz Cars. “The car is growing beyond its role as a mere means of transport and will ultimately become a mobile living space.”

Circular, but yes: "self-consciousness" && freedom.

What is "Autonomous" Technology?

Autonomy: Having the capability and freedom to self-direct to achieve mission objectives. An autonomous system makes choices and has the human's proxy for those decisions.




Challenges Associated with Autonomous Technology:

Technical	*Human/Autonomous System Interaction and Collaboration* *Scalable Teaming of Autonomous Systems* *Machine Reasoning and Intelligence* *Testing and Evaluation (T&E), Verification and Validation (V&V)*
Social	*Human-machine teaming* *Public perception of unmanned vehicles (land, sea, air)*
Economic	*Potential game-changing opportunity for many industries, including transportation, healthcare, security*

Not done with zombie freedom/autonomy.

But it's going to work like zombie *akrasia* …

Informal Definition of (Zombie) Akrasia

An action α_f is (Augustinian) akratic for an agent A at t_{α_f} iff the following eight conditions hold:

(1) A believes that A ought to do α_o at t_{α_o};

(2) A desires to do α_f at t_{α_f};

(3) A's doing α_f at t_{α_f} entails his not doing α_o at t_{α_o};

(4) A knows that doing α_f at t_{α_f} entails his not doing α_o at t_{α_o};

(5) At the time (t_{α_f}) of doing the forbidden α_f, A's desire to do α_f overrides A's belief that he ought to do α_o at t_{α_f}.

(6) A does the forbidden action α_f at t_{α_f};

(7) A's doing α_f results from A's desire to do α_f;

"Regret" (8) At some time t after t_{α_f}, A has the belief that A ought to have done α_o rather than α_f.

Cast in

$\mathcal{DCEC}^*$

becomes …

$$\mathsf{KB}_{rs} \cup \mathsf{KB}_{m_1} \cup \mathsf{KB}_{m_2} \ldots \mathsf{KB}_{m_n} \vdash$$

$$D_1 : \mathbf{B}(\mathsf{I}, \mathsf{now}, \mathbf{O}(\mathsf{I}^*, t_\alpha \Phi, happens(action(\mathsf{I}^*, \alpha), t_\alpha)))$$

$$D_2 : \mathbf{D}(\mathsf{I}, \mathsf{now}, holds(does(\mathsf{I}^*, \overline{\alpha}), t_{\overline{\alpha}}))$$

$$D_3 : happens(action(\mathsf{I}^*, \overline{\alpha}), t_{\overline{\alpha}}) \Rightarrow \neg happens(action(\mathsf{I}^*, \alpha), t_\alpha)$$

$$D_4 : \mathbf{K}\left(\mathsf{I}, \mathsf{now}, \begin{pmatrix} happens(action(\mathsf{I}^*, \overline{\alpha}), t_{\overline{\alpha}}) \Rightarrow \\ \neg happens(action(\mathsf{I}^*, \alpha), t_\alpha) \end{pmatrix}\right)$$

$$D_5 : \begin{array}{l} \mathbf{I}(\mathsf{I}, t_\alpha, happens(action(\mathsf{I}^*, \overline{\alpha}), t_{\overline{\alpha}}) \wedge \\ \neg \mathbf{I}(\mathsf{I}, t_\alpha, happens(action(\mathsf{I}^*, \alpha), t_\alpha) \end{array}$$

$$D_6 : happens(action(\mathsf{I}^*, \overline{\alpha}), t_{\overline{\alpha}})$$

$$D_{7a} : \begin{array}{l} \Gamma \cup \{\mathbf{D}(\mathsf{I}, \mathsf{now}, holds(does(\mathsf{I}^*, \overline{\alpha}), t))\} \vdash \\ \quad happens(action(\mathsf{I}^*, \overline{\alpha}), t_\alpha) \end{array}$$

$$D_{7b} : \begin{array}{l} \Gamma - \{\mathbf{D}(\mathsf{I}, \mathsf{now}, holds(does(\mathsf{I}^*, \overline{\alpha}), t))\} \nvdash \\ \quad happens(action(\mathsf{I}^*, \overline{\alpha}), t_\alpha) \end{array}$$

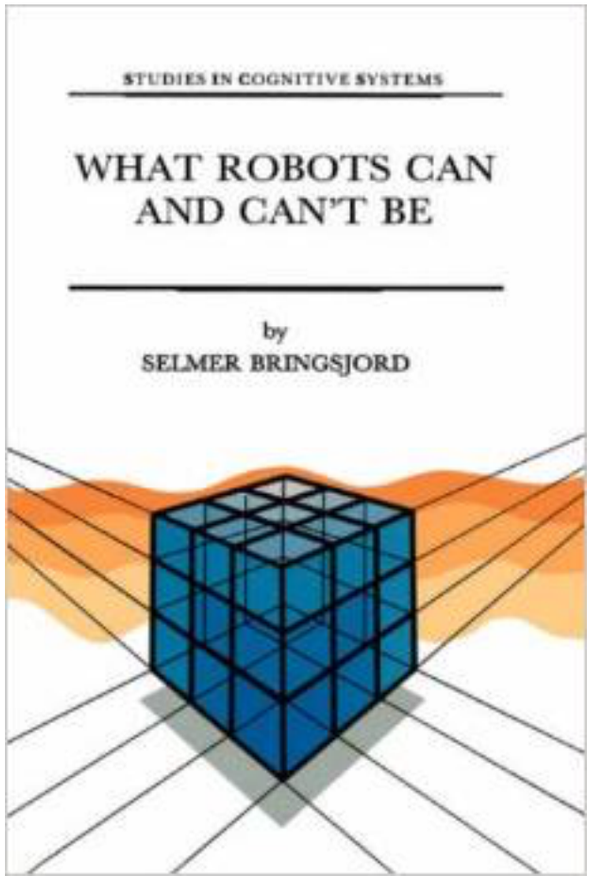
$$D_8 : \mathbf{B}\Big(\mathsf{I}, t_f, \mathbf{O}(\mathsf{I}^*, t_\alpha, \Phi, happens(action(\mathsf{I}^*, \alpha), t_\alpha))\Big)$$

Demos …

Need to do this for free will.

Has anyone provided a definition to exploit, in a manner analogous to what was done for Augustinian *akrasia*?

Actually, yes :).



Step 1 (Air Force):

$$Aut(m,t) \leftrightarrow \exists\, \alpha_1, \ldots, \alpha_m \exists\, t_1, t_2, \ldots, t_n \Diamond_{t^\star} (Freely\_Does(m, \alpha_i, t_j))$$

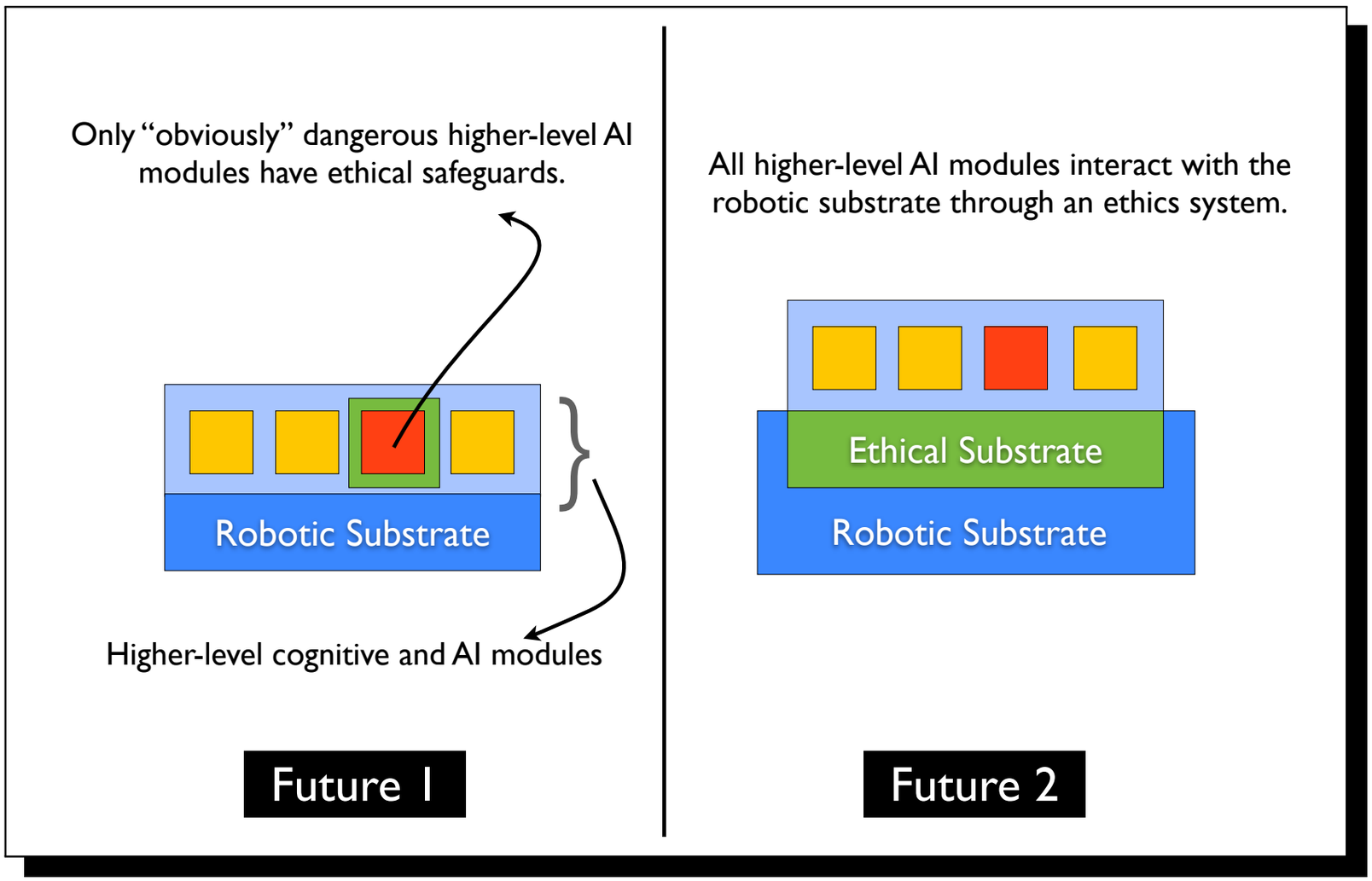
Step 2:

$$Freely\_Does(m, \alpha_i, t_j) \leftrightarrow$$

1. $Does(m, \alpha_i, t_j)$;
2. $\exists t(t < t_j \wedge Entertains(m, \alpha, t)$;
3. $\exists t(t < t_j \wedge Entertains(m, \bar{\alpha}, t)$;
4. $\exists t(t < t_j \wedge Wants(m, \alpha, t)$;
5. $\exists t(t < t_j \wedge Decides(m, t, \alpha, t_j)$;
6. $\Diamond_{t^\star} Does(m, \alpha_i, t_j)$;
7. $\Diamond_{t^\star} Does(m, \bar{\alpha}_i, t_j)$;

Philosophy & Logic For a Better Future!

Robots will have this — zombie — autonomy, but if as philosophical engineers we do our job well, even when these robots have motive, means, and opportunity, they won't form and execute destructive, unethical decisions—as long as protective engineering is installed at the op-sys level!



Naveen Sundar Govindarajulu and Selmer Bringsjord. "Ethical Regulation of Robots Must Be Embedded in Their Operating Systems" (book chapter, forthcoming), *A Construction Manual for Robot's Ethical Systems: Requirements, Methods, Implementations.*

slutten