الذكاء والقوة التحسينية والمزايا تم تقديم مجموعة كبيرة ومتنوعة من التعريفات لكلمة "الذكاء". والتعريف الذي يبدو أنه يلخص المضمون الأساسي في معظمها هو أن الذكاء يقيس قدرة العامل على تحقيق الأهداف في مجموعة واسعة من البيئات (ليغ وهوتر 2007). فالعقل لديه أهداف يحاول تحقيقها، والعقول الأكثر ذكاءً هي الأفضل في إيجاد وابتكار وتقييم الطرق المختلفة لتحقيق أهدافها. إن تعميم مفهوم الذكاء هو مفهوم القدرة على التحسين (Yudkowsky 2008a; Muehlhauser and Helm 2012)، وهو القدرة العامة للعقل على تحقيق أهدافه. في حين أن الذكاء مستمد من ما يعتبر عمومًا كليات "عقلية"، فإن قوة التحسين للوكيل هي أيضًا عامل من عوامل أخرى مثل حلفائه وموارده، وكذلك قدرته على الحصول على المزيد منها. إن الخطر الكبير الذي ينطوي عليه إنشاء عقول رقمية هو إمكانية خلق عقول تختلف أهدافها اختلافًا كبيرًا عن أهداف البشر، وينتهي بها الأمر بامتلاكها قوة تحسينية أكبر من تلك التي يمتلكها البشر. وموهلهاوزر وهيلم 2012). إذا تم إنشاء ذكاءات رقمية وانتهى بها الأمر بامتلاكها (كمجموعة) قوة تحسينية أكبر من تلك التي يمتلكها البشر، وكانت أهدافها مختلفة تمامًا عن أهداف البشر، فمن المرجح أن تعتبر العواقب سيئة للغاية من قبل معظم البشر. تحاول هذه الورقة البحثية تحليل عواقب إنشاء عقل رقمي من منظور القوة التحسينية التي قد تتراكم لديهم. وتسمى العوامل التي قد تؤدي إلى تراكم القوة التحسينية للعقول الرقمية أكثر من البشر بالمزايا،2. مزايا الأجهزة يمكن للعقل الرقمي الذي يعمل على نظام كمبيوتر ترقية النظام لاستخدام أجهزة أكثر قوة، في حين أن البشر البيولوجيين لا يستطيعون ترقية أدمغتهم بشكل كبير. لنفترض أن هناك حدًا أدنى من تكوين الأجهزة التي توفر للعقل الرقمي نفس قوة المعالجة والذاكرة تقريبًا مثل العقل البشري.2.1. إن مقدار قوة المعالجة الفائقة المطلوبة لتشغيل عقل رقمي غير معروف حالياً.1 بالنسبة للتحميلات، يضع ساندبرغ وبوستروم (2008) 1018 إلى 1025 فلوبس كأكثر كمية مطلوبة على الأرجح لتشغيل عقل رقمي في الوقت الحقيقي. ويقدّران أن الاتجاهات الحالية ستجعل هذه المستويات متاحة للشراء بتكلفة مليون دولار في الفترة من 2019 إلى 2044. مع الحفاظ على جميع التفاصيل ذات الصلة سليمة. وعلى النقيض من ذلك، فإن مصممي الذكاء الاصطناعي للذكاء الاصطناعي لديهم الحرية في استخدام أي خوارزمية عاملة، بغض النظر عن معقوليتها البيولوجية. لقد تطور الدماغ البشري ليعمل بطريقة تتناسب مع قيود علم الأحياء، والتي قد تكون مختلفة تمامًا عما يمكن أن يعمل بكفاءة على الكمبيوتر. وهذا يسمح باحتمال أن يتطلب الذكاء الاصطناعي للذكاء الاصطناعي المُعزَّز طاقة معالجة أقل بكثير من التحميل. تتراوح تقديرات كمية قوة المعالجة المطلوبة لتشغيل عقل ما بين أجهزة الكمبيوتر الحديثة (2.1.1،(Hall 2007. طاقة فائقة السرعة فالعقل الذي يجري تنفيذه على نظام ذي قوة تسلسلية أعلى بكثير يمكن أن يعمل على مقياس زمني أسرع مما نعمل نحن. على سبيل المثال، العقل الذي يتمتع بضعف القوة التسلسلية للدماغ البشري قد يختبر ما يعادل مرور ثانيتين مقابل كل ثانية تمر علينا، فيفكر بضعف كمية الأفكار في نفس الوقت. ستكون هذه الميزة ملحوظة بشكل خاص في اتخاذ القرارات الحرجة من حيث الوقت. حتى الميزة الصغيرة سوف تتراكم مع مرور الوقت الكافي. على مدار عام كامل، فإن فارق 10% في السرعة سيمنح العقل الأسرع أكثر من شهر إضافي. يودكوفسكي 1996؛ تشالمرز 2010)، يعمل الباحثون الرقميون الذين يعملون بسرعة متسارعة على تطوير أجهزة كمبيوتر أسرع. إذا تمكنت العقول التي تقوم بالبحث من الاستفادة من الأجهزة الأسرع التي أنتجتها، فإن الوقت اللازم لتطوير الجيل التالي من الأجهزة يمكن أن يستمر في التقلص لأن الباحثين سينجزون المزيد من العمل في نفس الوقت. أو إلى حاجز أساسي.2. زيادة الطاقة المتوازية وزيادة الذاكرة كانت التطورات الأخيرة في قوة الحوسبة متوازية بشكل متزايد بدلاً من التسلسلية. إذا استمر هذا الاتجاه، فقد لا يتمكن الحاسوب المستقبلي من استخدام قوة المعالجة الفائقة التي يتمتع بها لتحقيق زيادة مباشرة في السرعة مقارنةً بالعقل البشري، إذا كانت المهام المعنية لا يمكن موازاتها بشكل جيد. ينص قانون Amdahl على أنه إذا كان جزء f من أداء برنامج ما يمكن موازاته بجزء f من أداء البرنامج، فإن الزيادة في السرعة التي تمنحها n معالجات بدلاً من واحد هي 1/((1−f) +(f/n)) (Amdahl 1967). يشير Gustafson (1988) إلى أنه من الناحية العملية، فإن الجزء القابل للتوازي من المشكلة هو الذي ينمو مع إضافة البيانات، ويبقى الجزء المتسلسل ثابتًا. حتى لو لم تسمح زيادة عدد المعالجات بحل مشكلة ما في وقت أقل، فإنها يمكن أن تسمح بحل مشكلة أكبر بتفاصيل أفضل. نظرًا لأن الدماغ البشري يعمل بطريقة متوازية بشكل كبير، يجب أن تكون بعض الخوارزميات المتوازية للغاية على الأقل متضمنة في الذكاء العام. قد لا تسمح الطاقة الإضافية المتوازية الإضافية بتحسين السرعة بشكل مباشر، ولكن يمكن أن توفر شيئًا مثل ذاكرة عمل أكبر مكافئة. يمكن متابعة المزيد من مسارات التفكير في وقت واحد، ويمكن أخذ المزيد من الأشياء في الاعتبار عند التفكير في قرار ما. يبدو أن حجم الدماغ يرتبط بالذكاء داخل الفئران (أندرسون 1993)، وعبر الأنواع (دينر وآخرون 2007)، قد يقوم العقل الرقمي الذي لديه إمكانية الوصول إلى

شيفرته المصدرية بتعديل طريقة تفكيره مباشرة، أو إنشاء نسخة معدلة من نفسه. من أجل القيام بذلك، يجب أن يفهم العقل بنيته الخاصة بشكل جيد بما يكفي لمعرفة التعديلات المعقولة. يمكن بناء الذكاء الاصطناعي للذكاء الاصطناعي عن قصد بطريقة يسهل فهمها وتعديلها، بل قد يقرأ مستندات التصميم الخاصة به. قد تكون الأمور أصعب بالنسبة لعمليات التحميل، خاصة إذا لم يكن العقل البشري مفهوماً تماماً بحلول الوقت الذي يصبح فيه التحميل ممكناً. يمكن لأي نوع من أنواع العقل أن يختبر عدداً كبيراً من التدخلات الممكنة، فيخلق آلاف أو حتى ملايين النسخ من نفسه لمعرفة أنواع التأثيرات التي تحدثها التعديلات المختلفة. وفي حين أن بعض التعديلات يمكن أن تؤدي إلى مشاكل غير مرئية على المدى الطويل، يمكن حذف النسخ التي تحتوي على تعديلات ضارة أو محايدة، مما يفسح المجال لنسخ بديلة. (شولمان 2010). قد تنطوي المقاربات الأقل تجريبية على براهين رسمية لآثار التغييرات التي سيتم إجراؤها. Chalmers 2010) هو حالة يقوم فيها العقل بتعديل نفسه، مما يجعله قادرًا على تحسين نفسه بشكل أكبر. على سبيل المثال، وقد انخرط البشر، في عملية تحسين ذاتي متكررة، حيث أتاح تطوير تقنيات وأشكال جديدة من التنظيم الاجتماعي إمكانية التنظيم بشكل أفضل وتطوير تقنيات أكثر تقدمًا. ومع ذلك، ظل جوهر العقل البشري كما هو. والتي استمرت في إطلاق المزيد من التغييرات، فقد تكون النتيجة شكلاً محسناً بشكل كبير من الذكاء (1.1 .(Yudkowsky 2008a. الذكاء والقوة التحسينية والمزايا تم تقديم مجموعة كبيرة ومتنوعة من التعريفات لكلمة "الذكاء". والتعريف الذي يبدو أنه يلخص المضمون الأساسي في معظمها هو أن الذكاء يقيس قدرة العامل على تحقيق الأهداف في مجموعة واسعة من البيئات (ليغ وهوتر 2007). فالعقل لديه أهداف يحاول تحقيقها، والعقول الأكثر ذكاءً هي الأفضل في إيجاد وابتكار وتقييم الطرق المختلفة لتحقيق أهدافها. إن تعميم مفهوم الذكاء هو مفهوم القدرة على التحسين (Yudkowsky 2008a; Muehlhauser and Helm 2012)، وهو القدرة العامة للعقل على تحقيق أهدافه. في حين أن الذكاء مستمد من ما يعتبر عمومًا كليات "عقلية"، وكذلك قدرته على الحصول على المزيد منها. إن الخطر الكبير الذي ينطوي عليه إنشاء عقول رقمية هو إمكانية خلق عقول تختلف أهدافها اختلافًا كبيرًا عن أهداف البشر، وينتهي بها الأمر بامتلاكها قوة تحسينية أكبر من تلك التي يمتلكها البشر. يبدو أن التفضيلات والرغبات البشرية الحالية معقدة للغاية وغير مفهومة جيدًا: هناك احتمال قوي بأن مجموعة فرعية ضيقة جدًا من جميع الأهداف الممكنة ستؤدي في حال نجاحها إلى نتائج تعتبر مواتية للبشر (يودكوفسكي 2008 أ؛ وموهلهاوزر وهيلم 2012). إذا تم إنشاء ذكاءات رقمية وانتهى بها الأمر بامتلاكها (كمجموعة) قوة تحسينية أكبر من تلك التي يمتلكها البشر، فمن المرجح أن تعتبر العواقب سيئة للغاية من قبل معظم البشر. تحاول هذه الورقة البحثية تحليل عواقب إنشاء عقل رقمي من منظور القوة التحسينية التي قد تتراكم لديهم. وتسمى العوامل التي قد تؤدي إلى تراكم القوة التحسينية للعقول الرقمية أكثر من البشر بالمزايا، والعوامل التي قد تؤدي إلى تراكم قوة تحسينية أقل من البشر تسمى العيوب. يمكن للعقل الرقمي الذي يعمل على نظام كمبيوتر ترقية النظام لاستخدام أجهزة أكثر قوة، لنفترض أن هناك حدًا أدنى من تكوين الأجهزة التي توفر للعقل الرقمي نفس قوة المعالجة والذاكرة تقريبًا مثل العقل البشري. إن أي زيادة في موارد الأجهزة بعد هذه النقطة هي ميزة للأجهزة لصالح العقل الرقمي.2.1. قوة معالجة فائقة1 بالنسبة للتحميلات، يضع ساندبرغ وبوستروم (2008) 1018 إلى 1025 فلوبس كأكثر كمية مطلوبة على الأرجح لتشغيل عقل رقمي في الوقت الحقيقي. ويقدّران أن الاتجاهات الحالية ستجعل هذه المستويات متاحة للشراء بتكلفة مليون دولار في الفترة من 2019 إلى 2044. يحاول Anupload محاكاة وظيفة الدماغ البشري بأكملها بدقة، مع الحفاظ على جميع التفاصيل ذات الصلة سليمة. وعلى النقيض من ذلك، فإن مصممي الذكاء الاصطناعي للذكاء الاصطناعي لديهم الحرية في استخدام أي خوارزمية عاملة، بغض النظر عن معقوليتها البيولوجية. وهذا يسمح باحتمال أن يتطلب الذكاء الاصطناعي للذكاء الاصطناعي المُعزَّز طاقة معالجة أقل بكثير من التحميل. تتراوح تقديرات كمية قوة المعالجة المطلوبة لتشغيل عقل ما بين أجهزة الكمبيوتر الحديثة (Hall 2007)، و1014 فلوبس (2.1.1.(Bostrom 1997. طاقة فائقة السرعة يدرك البشر العالم على مقياس زمني مميز معين. فالعقل الذي يجري تنفيذه على نظام ذي قوة تسلسلية أعلى بكثير يمكن أن يعمل على مقياس زمني أسرع مما نعمل نحن. على سبيل المثال، العقل الذي يتمتع بضعف القوة التسلسلية للدماغ البشري قد يختبر ما يعادل مرور ثانيتين مقابل كل ثانية تمر علينا، فيفكر بضعف كمية الأفكار في نفس الوقت. على مدار عام كامل، فإن فارق 10% في السرعة سيمنح العقل الأسرع أكثر من شهر إضافي. وهذا من شأنه أن يسمح له بالتفوق على أي عقل بمهارات وموارد متساوية ولكن دون ميزة السرعة. في سيناريو "الانفجار السريع" (سولومونوف 1985؛ يودكوفسكي 1996؛ إذا تمكنت العقول التي تقوم بالبحث من الاستفادة من الأجهزة الأسرع التي أنتجتها، ويمكن أن يستمر ذلك إلى أن يتم الوصول إلى عنق زجاجة ما، مثل الوقت اللازم لبناء أجهزة الكمبيوتر فعليًا، أو إلى حاجز

أساسي.2.1.2. زيادة الطاقة المتوازية وزيادة الذاكرة كانت التطورات الأخيرة في قوة الحوسبة متوازية بشكل متزايد بدلاً من التسلسلية. إذا استمر هذا الاتجاه، إذا كانت المهام المعنية لا يمكن موازاتها بشكل جيد. ينص قانون Amdahl على أنه إذا كان جزء f من أداء برنامج ما يمكن موازاته بجزء f من أداء البرنامج، فإن الزيادة في السرعة التي تمنحها n معالجات بدلاً من واحد هي 1/((1–f) (Amdahl 1967) ((f/n +(f). قد يترجم الفرق بعدة أوامر من حيث الحجم في قوة الحوسبة إلى تغيير أكثر تواضعًا في السرعة. يشير Gustafson 1988) إلى أنه من الناحية العملية، فإن الجزء القابل للتوازي من المشكلة هو الذي ينمو مع إضافة البيانات، ويبقى الجزء المتسلسل ثابتًا. حتى لو لم تسمح زيادة عدد المعالجات بحل مشكلة ما في وقت أقل، فإنها يمكن أن تسمح بحل مشكلة أكبر بتفاصيل أفضل. نظرًا لأن الدماغ البشري يعمل بطريقة متوازية بشكل كبير، يجب أن تكون بعض الخوارزميات المتوازية للغاية على الأقل متضمنة في الذكاء العام. قد لا تسمح الطاقة الإضافية المتوازية الإضافية بتحسين السرعة بشكل مباشر، يمكن متابعة المزيد من مسارات التفكير في وقت واحد، والبشر (ماكدانييل 2005)، وعبر الأنواع (دينر وآخرون 2007)، مما يشير إلى أن زيادة القوة المتوازية يمكن أن تجعل العقل أكثر ذكاءً بشكل عام.3. مزايا التحسين الذاتي قد يقوم العقل الرقمي الذي لديه إمكانية الوصول إلى شيفرته المصدرية بتعديل طريقة تفكيره مباشرة، أو إنشاء نسخة معدلة من نفسه. من أجل القيام بذلك، يجب أن يفهم العقل بنيته الخاصة بشكل جيد بما يكفي لمعرفة التعديلات المعقولة. يمكن بناء الذكاء الاصطناعي للذكاء الاصطناعي عن قصد بطريقة يسهل فهمها وتعديلها، بل قد يقرأ مستندات التصميم الخاصة به. قد تكون الأمور أصعب بالنسبة لعمليات التحميل، يمكن لأي نوع من أنواع العقل أن يختبر عدداً كبيراً من التدخلات الممكنة، فيخلق آلاف أو حتى ملايين النسخ من نفسه لمعرفة أنواع التأثيرات التي تحدثها التعديلات المختلفة. وفي حين أن بعض التعديلات يمكن أن تؤدي إلى مشاكل غير مرئية على المدى الطويل، يمكن إخضاع كل نسخة لاختبارات مكثفة مختلفة على مدى فترة زمنية طويلة لتقدير آثار التعديلات. مما يفسح المجال لنسخ بديلة. (شولمان 2010). قد تنطوي المقاربات الأقل تجريبية على براهين رسمية لآثار التغييرات التي سيتم إجراؤها. على سبيل المثال، مما يسمح له بعد ذلك بملاحظة أوجه القصور في نفسه. إن تصحيح أوجه القصور هذه من شأنه أن يوفر وقت المعالجة ويسمح للذكاء الاصطناعي بملاحظة المزيد من الأشياء التي يمكن تحسينها. وقد انخرط البشر، إلى حد محدود،